Université Assane Seck de Ziguinchor
UFR Sciences et Technologies
Département Informatique

Mémoire de fin d'études

Pour l'obtention du diplôme de Master
Mention : Informatique
Spécialité : Génie Logiciel

Sujet :

Extraction d'entités nommées spécifiques à partir de données biomédicales non structurées

Présenté et soutenu par :
M. Ibrahima NDAO

Sous la direction de :
Dr Khadim DRAME

Sous la supervision de :
Pr Youssou FAYE

Membres du jury

Pr. Youssou FAYE	Professeur assimilé	Président	UASZ
Dr. Khadim DRAME	Maître de conférences titulaire	Encadrant	UASZ
Dr. Gorgoumack SAMBE	Maître de conférences assimilé	Examinateur	UASZ
Dr. Ibrahima DIOP	Maître de conférences titulaire	Rapporteur	UASZ

Année universitaire 2020/2021

Remerciements

« J'entends et j'oublie. Je vois et je me souviens. Je fais et je comprends. » Confucius (551 à 479 av. J.-C.)

Tout d'abord, je remercie **Allah** le **Tout-Puissant** de m'avoir donné la force, la patience et le courage d'accomplir ce travail.

Mes chaleureux remerciements à mon encadrant **Monsieur Khadim DRAME**, pour son soutien, sa disponibilité, ses encouragements et les nombreuses discussions qui m'ont permis d'y voir plus clair.

Mes vifs remerciements aux membres du jury, **Monsieur Youssou FAYE**, **Monsieur Gorgoumack SAMBE** et **Monsieur Ibrahima DIOP**, enseignants-chercheurs à l'université Assane Seck de Ziguinchor, pour l'intérêt qu'ils ont porté à mon travail, le temps qu'ils ont bien voulu consacrer à l'évaluation de ce mémoire et de l'enrichir par leurs propositions.

J'adresse mes plus sincères remerciements à mes parents, mes frères et mes sœurs qui m'ont soutenu et encouragé durant tout mon cursus.

Un grand merci à Amadou Diouhé Diallo, Dr Lamine Faty, Mactar Gane, Cédrique Nyafouna et Bacary Sanokho pour vos corrections.

Enfin, je remercie mes amis, mes camarades de promos, mes camarades de chambres et à toute personne qui a contribué de près ou de loin à l'élaboration de ce travail.

Dédicaces

A mon père Abdoulaye NDAO, ma mère Thioro BA,

Mes frères et sœurs,

Mes neveux et nièces

Mon défunt camarade de classe Mamour DIOUF.

Résumé

Les technologies de l'information et de la communication sont communément utilisées dans des activités quotidiennes. Une grande quantité d'informations est générée via les réseaux sociaux, les blogs, les forums, etc. Le traitement de cette masse d'informations, représentée souvent dans des formats non structurés (textes) ou semi structurés, de façon robuste et performante devient un grand enjeu. Le traitement automatique des langues (TAL) s'intéresse à cette question et il fournit des outils pour réaliser cette tâche. L'extraction d'informations, qui est un sous domaine du TAL, vise à extraire des informations pertinentes à partir des textes. L'extraction d'entités nommées qui peut être vue comme une sous tâche de l'extraction d'informations, suscite un grand engouement aujourd'hui. Différentes approches sont proposées dans la littérature : approches symboliques basées sur des lexiques et des dictionnaires, approches statistiques basées sur l'apprentissage automatique (machine learning, ML en anglais) et les approches hybrides. Les outils utilisant ces approches ont connu un grand succès notamment sur des textes en anglais (avec des précisions et f-mesures de plus de 90%), mais ils restent moins performants pour les langues comme le français.

Ce mémoire présente dans un premier temps un état de l'art sur l'extraction d'entités nommées dans le domaine général ainsi que dans le domaine biomédical : son historique, les approches utilisées et les outils existants. Dans un second temps, nous allons décrire le modèle proposé pour la reconnaissance d'entités nommées sur des données biomédicales en utilisant une approche à base de règles exploitant des lexiques, des dictionnaires et des règles. L'évaluation de notre approche sur des données standards a donné des résultats satisfaisants (avec une précision et un rappel de plus de 94%).

Mots clés : Reconnaissance d'entités nommées(REN), Traitement automatique des langues(TAL), Apprentissage automatique, Apprentissage profond, domaine biomédical.

Abstract

Information and communication technologies are commonly used in everyday activities. A large amount of information is generated via social networks, blogs, forums, etc. Processing this mass of information, often represented in unstructured (text) or semi-structured formats, in a robust and efficient way is becoming a major challenge. Natural language processing (NLP) is addressing this issue and providing tools to perform this task. Information extraction, which is a sub-domain of NLP, aims at extracting relevant information from texts. Named entity extraction, which can be seen as a subtask of information extraction, is currently very popular. Different approaches are proposed in the literature: symbolic approaches based on lexicons and dictionaries, statistical approaches based on machine learning (ML) and hybrid approaches. Tools using these approaches have been very successful, especially on English texts (with accuracies and f-measurements of more than 90%), but they remain less efficient for languages like French.

Firstly this paper presents a state of the art on named entity extraction in the general domain as well as in the biomedical domain: its history, the approaches used and the existing tools. Then, we will describe the proposed model for named entity recognition on biomedical data using a rule-based approach exploiting lexicons, dictionaries and rules. The evaluation of our approach on standard data gave satisfactory results (with precision and recall of more than 94%).

Keywords: Named Entity Recognition (NER), Natural Language Processing (NLP), Machine Learning, Deep Learning, Biomedical domain.

Table des matières

Liste des illustrations

Liste des formules

Liste des tableaux

Liste des abréviations

BERT	Bidirectional Encoder Representations from Transformers
BioNER	Biomedical Named Entity Recognition
BioNLP	Biomedical Natural Language Processing
CNN	Convolutional Neural Networks
CoNLL	Conference on Natural Language Learning
CRF	Conditional Random Field
CSV	Comma Separated Values
DARPA	Defense Advanced Research Projects Agency
DEFT	Défi de Fouille de Textes
DL	Deep Learning
EEN	Extraction d'Entités Nommées
EI	Extraction d'Informations
EN	Entité Nommée
ENs	Entités Nommées
ESTER	Evaluation des Systèmes de Transcription enrichie d'Emissions Radiophoniques
ETAPE	Evaluation en Traitement Automatique de la Parole
ETER	Entity Tree Error Rate
HMM	Hidden Markov Model
MEMM	Modèle de Markov à Entropie Maximale
MET	Multilingual Entity Task
ML	Machine Learning
MUC	Message Understanding Conference
NER	Named Entity Recognition
NLP	Natural Language Processing
OCR	Optical Character Recognition
POS	Part-Of-Speech
REN	Reconnaissance d'Entités Nommées
RNN	Recurrent Neural Networks
ROC	Reconnaissance Optique de Caractères
SER	Slot Error Rate
SVM	Support Vector Machine

TAL	Traitement Automatique des Langues
TALN	Traitement Automatique du Langage Naturel
UIMA	Unstructured Information Management Architecture
UMLS	Unified Medical Language System

Introduction générale

A l'heure actuelle, l'utilisation des outils informatiques est devenue incontournable dans quasiment tous les domaines. On note qu'au quotidien de grandes quantités de données sont produites et partagées à travers les médias sociaux, les forums, les sites d'informations, etc. Le traitement de ces grandes masses de données, de par leur nature hétérogène (sons, vidéos, textes, images, etc.), a motivé plusieurs travaux de recherche surtout dans le domaine du traitement automatique des langues (TAL) ou traitement automatique du langage naturel (TALN) (ou natural language processing (NLP[1]) en anglais). Ces données, plus particulièrement les données textuelles, sont représentées dans des formats qui ne sont pas normalisés[2]. Ce qui rend leurs exploitations de plus en plus difficiles.

L'extraction d'entités nommées est utilisée dans beaucoup d'applications du TAL comme la traduction automatique, la veille technologique[1], la classification de documents, l'extraction des relations entre entités[2][3], l'extraction d'évènements[4], la fouille d'opinions[5], etc. Ainsi, la création d'outils permettant d'extraire efficacement les informations pertinentes à partir de corpus de textes devient incontournable. Néanmoins, ces outils doivent faire face à de nombreuses difficultés :

- les données textuelles sont constituées d'informations non structurées ;
- le langage naturel présente un caractère imprévisible (langage familier, faute d'orthographe, mauvais usage de la majuscule, etc.) ;
- les ressources nécessaires ne sont pas toujours disponibles dans certaines langues comme le français.

Dans ce mémoire, nous nous intéressons à l'extraction d'entités nommées spécifiques (âge, genre, issue, origine d'admission) dans des corpus de textes biomédicaux (cas cliniques) en français. Dans un premier temps, nous avons fait un état de l'art de l'extraction d'entités nommées dans le domaine général et dans le domaine biomédical. Ensuite, nous avons développé un système d'extraction d'entités nommées de types « âge » et « genre » à partir de données de cas cliniques en français.

[1] NLP (Natural Language Processing) ou TAL (Traitement automatique des langues) est un des domaines de l'intelligence artificielle les plus prolifiques, il consiste à comprendre et à traiter le langage humain.

[2] Une donnée non structurée est une donnée qui est manipulée (stockée) dans son format initial et est non traitée avant son utilisation (Exemple : mails, tweets…).

Une entité nommée (EN) est une unité textuelle (mot ou groupe de mots) qui fait référence à une entité unique et concrète appartenant à un domaine spécifique (social, économique, agronomique, géographique, biomédical, etc.) [6].

Le terme « entité nommée » désigne principalement les noms de personnes, d'organisations et de lieux. Cette notion peut s'étendre sur d'autres catégories comme les expressions temporelles, les expressions numériques, les évènements, etc. Dans les domaines d'applications comme le domaine biomédical, les entités nommées (ENs) désignent entre autres les noms de maladies, de gènes, de médicaments, de protéines, etc.

Ce mémoire comprend quatre chapitres :

- le premier chapitre fait une revue de la littérature sur la reconnaissance (l'extraction) d'entités nommées (REN) dans le domaine général. Il comporte trois sections. La première section présente l'historique de la tâche d'extraction d'entités nommées et sa définition, les différentes formes et catégories d'ENs et enfin les différentes phases de traitement pour la REN. La deuxième section décrit les différentes étapes et les problématiques liées à l'extraction d'ENs. La troisième section présente les différents domaines d'application de la REN.
- le deuxième chapitre comporte trois sections. La première section fait un état de l'art sur les différentes approches de la REN, leurs avantages et leurs limites. La deuxième section décrit les métriques qui permettent d'évaluer un système de REN. Enfin, dans la troisième section, quelques outils d'extraction d'ENs existants sont exposés.
- Le troisième chapitre porte sur la reconnaissance d'entités nommées (REN) dans le domaine biomédical. Il est constitué de trois sections. Dans la première section, nous nous intéressons à la notion d'entité nommée et aux différents types d'entités nommées dans le domaine biomédical. La deuxième section expose les défis et les problématiques sur l'extraction d'ENs dans le domaine biomédical. Enfin, dans la troisième section, quelques outils et ressources existants dans le domaine biomédical sont décrits.
- Le quatrième et dernier chapitre présente notre approche pour la création d'un modèle d'extraction d'entités nommées dans le domaine biomédical. Il comprend deux sections. La première section présente la méthodologie adoptée pour la création du modèle. Et dans la seconde section, le corpus, les métriques, les outils utilisés, les configurations, les résultats et une discussion sur le modèle proposé sont décrits.

Chapitre 1 : Reconnaissance d'entités nommées dans le domaine général

Introduction

Une grande masse de données textuelles est disponible sur les plateformes numériques. Ainsi, l'exploitation automatiquement de ces textes devient un besoin capital. Les recherches menées pour le développement d'outils permettant une compréhension automatique de ces données sont en nette croissance. Elles s'appuient sur les méthodes et outils du traitement automatique des langues (TAL) comme l'extraction d'entités nommées ou la reconnaissance d'entités nommées (REN, named entity recognition, NER en anglais). La reconnaissance d'entités nommées (REN) consiste à extraire des éléments pertinents à partir de textes pour en faire une collection de données bien ordonnées.

Dans ce chapitre, nous allons d'abord rappeler l'historique de la tâche d'extraction d'entités nommées et présenter les différentes définitions proposées dans la littérature. Ensuite, les différentes formes, catégories et phases de la reconnaissance d'entités nommées (REN) seront décrites. Enfin, nous mettrons en exergue les différentes étapes et les problématiques liées à la REN ainsi que ses domaines d'application.

I.1. Reconnaissance d'entités nommées (REN)

I.1.1. Historique

Selon Maud Ehrmann [7], la tâche d'extraction d'entités nommées est née à l'issue de l'expansion de la tâche d'extraction d'informations. Celle-ci est définie par Thierry Poibeau [8] comme suit : *« L'extraction d'informations est une tâche qui consiste à remplir automatiquement une banque de données à partir de textes écrits en langue naturelle ».* L'extraction d'entités nommées quant à elle consiste à extraire des mots ou des groupes de mots correspondants à des entités catégorisables comme des noms propres (noms de personnes, de lieux et d'organisations) ou des descriptions définies (dates, nombres, etc.). C'est lors d'une série de sept conférences appelée les conférences MUC[3] (Message Understanding Conference**,** Conférence sur la compréhension de messages en français) [9] [10] qui se sont tenues entre 1987 et 1998 [8] [11] et organisées par différentes institutions américaines et financées par la

[3] https://www-nlpir.nist.gov/related_projects/muc/

DARPA[4] aux Etats-Unis [7] [12]. Ces conférences [13] avaient pour objectif de motiver la recherche sur la compréhension automatique des messages militaires en extrayant les noms de personnes, de lieux, d'organisations et entre autres dans l'unique but de répondre à la question suivante : *« De quoi parle le texte ? »* [14].

Ainsi, c'est à partir de la 6^{iem} conférence (MUC 6[5] en 1996) [15] que les chercheurs ont initié la tâche d'extraction d'entités nommées (noms de personnes, d'organisations, de lieux, expressions numériques, d'expressions temporelles, etc.). Elle aura pour unique but d'extraire les unités lexicales les plus appropriées du texte et d'en faire une description ordonnée (structurée). Selon Grouin [16], cette tâche vise *« à accéder aux informations contenues dans des textes dans la perspective de répondre à des questions basiques comme Qui ? Quoi ? Quand ? Où ? Comment ? Pourquoi ? ».*

L'ensemble des traitements effectués lors de la conférence MUC 6 portait sur la langue anglaise et des messages militaires. Les types d'entités extraites étaient de types ENAMEX, NUMEX, TIMEX (cf. section I.1.3).

Il s'en est suivi d'autres campagnes d'évaluations [17] de traitement automatique des langues traitant la tâche d'extraction d'entités nommées qui diffèrent par les types d'entités extraites ou bien de la langue concernée. Parmi ces campagnes, on peut en citer notamment celle des conférences CoNLL (Conference on Natural Language Learning) pour l'anglais, l'espagnol et l'allemand [18], ESTER[6] (Evaluation des Systèmes de Transcription enrichie d'Emissions Radiophoniques) et ETAPE[7] (Evaluation en Traitement Automatique de la Parole) pour le français, MET[8] (Multilingual Entity Task) pour le chinois et le japonais.

I.1.2. Définitions

Ces dernières années, la reconnaissance d'entités nommées (REN) a fait l'objet de plusieurs recherches avec un champ d'application qui est multiple et varié. Néanmoins, il n y a toujours pas eu de consensus sur la définition d'entité nommée (EN) [19] [20] et jusque-là, on note des tentatives de définitions et de redéfinitions de cette notion par les chercheurs [4].

[4] https://www.darpa.mil/
[5] https://cs.nyu.edu/~grishman/muc6.html
[6] http://www.afcp-parole.org/campagne-devaluation-ester/
[7] http://www.afcp-parole.org/campagne-devaluation-etape/
[8] https://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_proceedings/overview.html

La définition d'une entité nommée a beaucoup évolué dans le temps. En effet, elle a été proposée lors de la campagne d'évaluation MUC 6 [12] durant laquelle les entités nommées désignaient des termes qui font référence aux noms propres tels que les noms de personnes, de lieux et d'organisations [6] [17]. Par la suite, lors de la conférence MUC 7, cette définition est étendue: « Une entité nommée désigne les noms de personnes, de lieux, d'organisations, d'expressions numériques et temporelles ». Ces définitions, proposées lors de ces conférences, sont basées particulièrement sur une démarche énumérative des entités nommées par rapport aux besoins applicatifs.

Par ailleurs, d'autres définitions ont été ensuite proposées par des chercheurs qui se basent pour la plupart sur des critères de référence ou d'unicité. Par exemple, Fotsoh propose la définition suivante [19] : *« Une EN peut être définie comme une unité linguistique (syntagme), identifiable de façon unique dans un contexte précis et qui renvoie à un objet du monde réel. »*

Le Meur définit les ENs dans le domaine général comme suit : *« Les ENs sont des types d'unités lexicales particulières qui font référence à une entité du monde concret dans certains domaines spécifiques notamment humains, sociaux, politiques, économiques, ou géographiques et qui ont un nom.»* [12].

Ehrmann quant à lui définit une EN comme suit [7]: *« Etant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus. »*

Tandis que, selon Fotsoh Tawofaing, la définition la plus complète est celle proposée dans l'encyclopédie Atalapédia de l'ATALA[9] (Association pour le Traitement Automatique des Langues) [19].D'ailleurs, celle-ci a été reprise par Maud Ehrmann dans sa thèse [7]:

« Les entités nommées désignent l'ensemble des noms de personnes, de lieux, d'entreprises, etc… contenues dans un texte. On ajoute souvent à ces éléments les dates et d'autres données chiffrées. Par extension, les entités désignent parfois les éléments de base pour une tâche donnée (par exemple, les noms de gènes dans le cadre de l'étude des textes de biologies) […] Ces séquences référentielles sont primordiales pour beaucoup d'applications linguistiques, que ce soit la recherche ou l'extraction d'information, la traduction automatique ou la compréhension de textes. »

[9] http://www.atala.org/

Bien que toutes ces définitions soient pour la plupart émises en fonction des notions linguistiques et applicatives. Damien Nouvel, quant à lui, voit que la notion d' « *entité nommée* » est issue d'un modèle applicatif. Donc une définition aussi générale de cette notion ne peut dépendre d'un corpus ou d'un modèle [21]. Alors, il propose la définition suivante : *« Les entités nommées désignent des objets mentaux de manière stable, à partir desquels il est attendu qu'une représentation logique opère»*

Cependant, on peut dire qu'une entité nommée peut être définie comme toute information ou connaissance du texte qui correspond à un objet perceptible du monde réel appartenant à un domaine spécifique (économique, géographique, médical, etc.) permettant une meilleure compréhension du texte.

Le tableau suivant illustre des exemples d'entités nommées :

Entités nommées
Barack Obama
Amérique
18 Avril 2007
UNESCO

Tableau 1: Exemples d'entités nommées

Depuis la conférence MUC 6, les entités nommées ont été classées dans différentes catégories selon les types d'entités.

I.1.3. Typologies d'entités nommées

Il existe plusieurs types d'entités nommées. De plus, plusieurs campagnes d'évaluations existent [4] qui pour chacune d'elles a proposé une typologie spécifique dépendant fortement de la langue, du domaine, de l'auteur, etc. En dépit de cela, la typologie la plus répandue reste celle proposée lors de la conférence MUC 6 où les entités nommées extraites étaient au nombre de sept types et classées en trois catégories (classes) distinctes [15]:

- la classe **ENAMEX**[10] : cette classe renferme les types d'entités nommées décrivant les noms propres (personne, lieu, organisation) ;

[10] Entity NAMe EXpression

- la classe **TIMEX**[11] : cette classe renferme les expressions temporelles et les dates ;
- la classe **NUMEX**[12] : elle comprend les valeurs monétaires et les pourcentages.

Le tableau ci-après présente des exemples d'entités nommées avec leurs types respectifs :

Entités nommées	**Types**
Barack Obama	Nom de personne **(PERS)**
Amérique	Nom de lieu **(LOC)**
18 Avril 2007	Expression temporelle **(DATE)**
UNESCO	Nom d'organisation **(ORG)**

Tableau 2 : Exemples d'entités nommées avec leurs types respectifs

L'illustration suivante montre un exemple de typologie présenté lors de la campagne d'évaluation MUC 6 avec leurs catégories distinctes :

Mr. <ENAMEX TYPE=« PERSON » > Dooner </ENAMEX> met with <ENAMEX TYPE=« PERSON » > Martin Puris </ENAMEX>, president and chief executive officer of <ENAMEX TYPE=« ORGANIZATION » > Ammirati & Puris </ENAMEX>, about <ENAMEX TYPE=« ORGANIZATION » > McCann </ENAMEX>'s acquiring the agency with billings of <NUMEX TYPE=« MONEY » > $400 million </NUMEX>, but nothing has materialized.

Illustration 1 : Exemple d'annotation d'entités nommées MUC 6 [7]

Toutefois, les classes ou types d'entités à extraire dépendent entièrement du type d'application (du domaine dans lequel on extrait les entités). Par exemple, dans le domaine :

- biomédical : noms de médicaments, de maladies, de gènes, de protéines, etc. ;
- agronomique : types de sols, de climats, de semences, d'engrais, etc.
- géographique [22]: noms de région, de pays, etc.

Cependant, les entités nommées peuvent se présenter dans le texte sous plusieurs formes.

[11] TIMe EXpression
[12] NUMeric EXpression

I.1.4. Formes d'entités nommées

Les entités nommées sont des unités lexicales représentées sous forme de mots ou de groupe de mots. Dans le texte, les entités nommées se présentent sous plusieurs formes qui peuvent être classées en deux catégories distinctes [23]:

- **Entité nommée simple :** Une entité nommée est dite simple, si celle-ci est composée d'un seul terme (token ou mot).

  Exemple :

  - ✓ Nom de personne : « Abdou », « Michel », etc.
  - ✓ Lieu : « Sénégal », « Mauritanie », etc.
  - ✓ Organisation : « ONU », « Google », etc.
- **Entité nommée complexe** : Une entité nommée complexe quant à elle est une entité nommée constituée de plusieurs termes.

  Exemple :

  - ✓ Nom de personne : « Marcel Diop », « Aminata Ndiaye », etc.
  - ✓ Lieu : « Afrique du sud », « Etas Unis d'Amérique », etc.

I.1.5. Phases de la reconnaissance d'entités nommées (REN)

La reconnaissance d'entités nommées (REN) est un sous domaine de l'extraction d'informations qui consiste à extraire des éléments pertinents à partir de corpus de textes. Elle est constituée d'un ensemble d'étapes qui permettent de déterminer dans le texte, les entités nommées du domaine et de les catégoriser dans leurs classes respectives [24]. Cette tâche s'effectue en deux étapes majeures qui sont étroitement liées : la détection et la catégorisation. Dans un système de reconnaissance d'entités nommées, ces deux étapes peuvent s'effectuer de manière séquentielle ou simultanée.

- **La détection :** c'est la phase qui consiste à identifier dans le texte simultanément les frontières (les bornes exactes des entités) entre les éléments qui sont potentiellement des entités nommées du domaine concerné [25].

L'exemple suivant montre une illustration de cette phase :

Le maréchal de Biron, qui n'épiait qu'une telle occasion, en étant averti, feint de venir avec son armée près de là pour joindre à un passage de rivière Monsieur de Cornusson, sénéchal de Toulouse, qui lui amenait des troupes, et au lieu d'aller là, tourne vers Nérac, et sur les neuf heures du matin s'y présente avec toute son armée en bataille, prêt et à la volée du canon.

Illustration 2: Exemple d'identification des limites des entités nommées [25]

- **La catégorisation :** La catégorisation quant à elle, consiste à attribuer la typologie (catégorie) spécifique à des entités déjà identifiées durant la phase de détection [25]. Elle est illustrée par l'exemple suivant.

Le maréchal de Biron[pers], qui n'épiait qu'une telle occasion, en étant averti, feint de venir avec son armée près de là pour joindre à un passage de rivière Monsieur de Cornusson[pers], sénéchal de Toulouse[pers], qui lui amenait des troupes, et au lieu d'aller là, tourne vers Nérac[loc], et sur les neuf heures du matin[temp] s'y présente avec toute son armée en bataille, prêt et à la volée du canon.

Illustration 3 : Exemple de catégorisation d'entités nommées [25]

L'identification et la catégorisation d'entités nommées dans du texte s'appuient sur deux indices définis par McDonald [26]: l'indice interne qui permet de reconnaitre une entité nommée à travers sa forme et l'indice externe qui utilise les contextes à gauche et à droite pour déterminer et bien catégoriser les entités nommées.

Les indices internes permettent, à partir de la forme du mot ou du groupe de mots dans le texte, de déterminer si celui-ci est une entité nommée ou non. Parmi les indices internes, on peut avoir :

- **les informations graphiques** [26]**:** à travers la représentation graphique du mot ou du groupe de mots dans le texte, on peut s'apercevoir que cet élément peut éventuellement être une entité nommée du domaine. Par exemples :
  - un mot commençant par une majuscule ;
  - un mot écrit totalement en majuscule ;
  - un sigle ou une abréviation, etc.

  Exemple : « Obama est reçu par le secrétaire général de l'ONU.»

Dans cet exemple, les mots « **Obama** » et « **ONU** » sont succinctement reconnus comme des entités nommées.

- **les informations concernant la ponctuation, les caractères spéciaux et numériques** [26]**:** cet indice permet, à partir des caractéristiques morphologiques du mot, d'identifier et de classer les entités nommées. Il prend en compte les ponctuations, les caractères spéciaux et les expressions numériques présents dans les mots afin de les catégoriser. Il permet de reconnaitre souvent les entités nommées de types organisations, les dates, les pourcentages, etc.

  Exemples :

  - un mot terminant par « .inc » montre la présence d'une entité de type entreprise (organisation) ;
  - « 17/12/1985 », « 1990 », « 12 Avril 2002 », les termes écrits sous cette forme spécifient généralement des entités de type date.
- **les informations morpho-syntaxiques** [26]**:** l'utilisation d'un étiqueteur morpho-syntaxique permet d'associer à chaque token du texte une catégorie grammaticale (verbe, nom propre, conjonction, etc.). Ces informations peuvent servir à catégoriser une entité nommée. Par exemple, un mot étiqueté comme un nom propre est fort probable qu'il soit ou fasse partie d'une entité.

Les indices externes servent à lever l'ambiguïté sur les indices internes pour ne pas aboutir à une erreur de catégorisation [26]. Différents types d'informations sont utilisés :

- **la position du mot dans la phrase :** cet indice permet de lever l'ambiguïté surtout sur le cas d'un mot commençant par une majuscule. Puisqu'une phrase commence toujours par une majuscule, le premier mot d'une phrase bien qu'il débute par une majuscule ne garantit pas forcément la présence d'une entité nommée tandis qu'un mot commençant par une majuscule et n'étant pas en début de phrase lui a de forte chance d'être une entité nommée.
- **les informations concernant les autres occurrences de l'entité nommée potentielle dans le document ou dans le corpus** : cet indice traite le cas de la catégorisation des mots polysémiques (mots ayant plusieurs sens). Par exemple, si dans un contexte le mot « orange » présente une ambiguïté et qu'il ait une même occurrence du mot qui a été

catégorisée comme une couleur, alors il peut aussi être considéré comme entité nommée de type couleur.

I.2. Etapes et problématiques de la REN

I.2.1. Etapes pour l'extraction d'entités nommées

Les données textuelles, du fait de leur nature non structurée ou semi structurée, sont pour la plupart représentées sous forme de long texte constitué de phrases. Alors qu'une entité nommée correspond à un mot ou à un groupe de mots de ces textes. Ainsi, l'extraction des unités textuelles nécessite un ensemble de prétraitements allant du découpage du texte en phrases puis en tokens (mots) jusqu'à la reconnaissance des mots ou groupes de mots éligibles comme entités nommées du domaine. Ce traitement effectué de façon séquentielle constitue les différentes étapes pour l'extraction des entités nommées dans des textes.

La figure suivante présente les différentes étapes de ce processus.

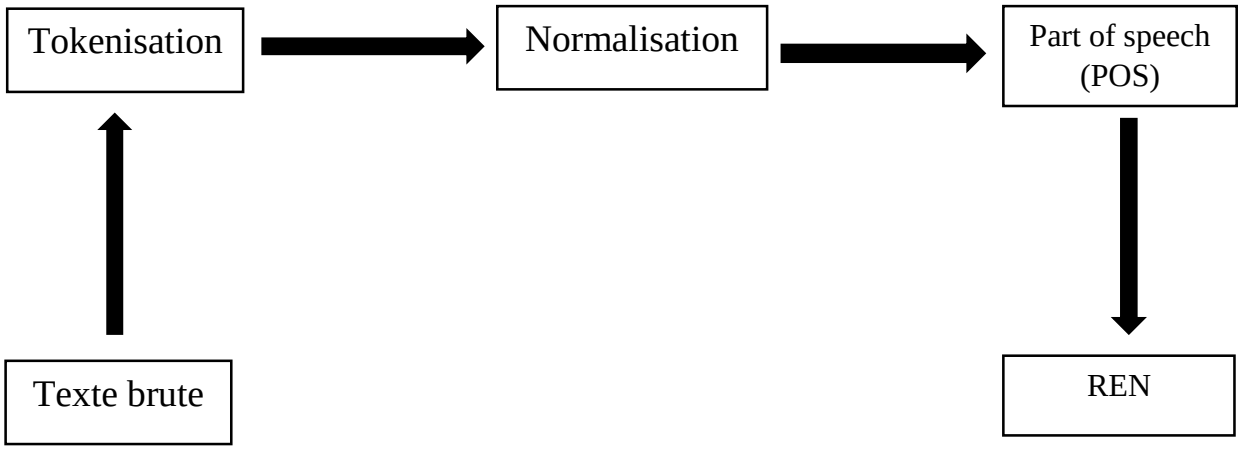

Illustration 4 : Processus d'extraction des entités nommées

- **La tokenisation** [14]**:** c'est l'étape qui consiste à découper le texte en termes ou tokens (mots, ponctuations, etc.).

> **Exemple : « Souleymane Diop, 24 ans travaille à la SONATEL»**
>
> Apres tokenisation, on obtient les tokens suivants:
>
> « Souleymane », « Diop », «,», « 24 », « ans », « travaille », « à », « la », « SONATEL »

Illustration 5 : Exemple de tokenisation d'une phrase

- **La normalisation :** Elle permet de normaliser l'écriture d'un mot par un de ses dérivés. En TAL, la normalisation des mots est effectuée soit par la lemmatisation (lemmatization en anglais), soit par la racinisation (stemming en anglais).
    - **La lemmatisation** [14]**:** Elle consiste à déterminer la forme canonique de chaque terme (token) du texte. Elle représente les mots au pluriel par leurs correspondances au masculin et au singulier et les verbes par leurs infinitifs.

Exemple : « Souleymane Diop, 24 ans travaille à la SONATEL»

Token	Lemme
Souleymane	Souleymane
Diop	Diop
,	,
24	24
Ans	An
Travaille	Travailler
A	A
La	La
SONATEL	SONATEL

Tableau 3 : Exemple de lemmatisation d'une phrase

- **la racinisation** : Elle consiste à déterminer la racine (radical) de chaque token (mot ou jeton). Cette racine n'est pas forcément un mot valide.

Exemple : « Souleymane Diop, 24 ans travaille à la SONATEL»

Token	Racine (stem)
Souleymane	Souleyman
Diop	Diop
,	,
24	24
Ans	An
Travaille	Travail
A	A
la	La
SONATEL	SONATEL

Tableau 4 : Exemple de racinisation d'une phrase

- **L'étiquetage morpho-syntaxique (part-of-speech (POS))** [14]**:** Elle consiste à donner la nature grammaticale (nom, adjectif, adverbe, verbe, pronom, déterminant, etc.) des différents tokens recensés.

Exemple : « Souleymane Diop, 24 ans travaille à la SONATEL»

Token	POS tagging
Souleymane	PROPN
Diop	PROPN
,	PUNCT
24	NUM
Ans	NOUN
Travaille	VERB
à	ADP
La	DET
SONATEL	NOUN

Tableau 5 : Exemple d'étiquetage morpho-syntaxique d'une phrase

- **La REN :** Cette étape, qui est la phase finale du processus, permet d'attribuer à chaque entité identifiée une catégorie spécifique.

Exemple : « Souleymane Diop, 24 ans travaille à la SONATEL»

Token	REN
Souleymane Diop	PER
SONATEL	ORG

Tableau 6 : Exemple de REN dans une phrase

Il y a d'autres étapes qui peuvent participer au processus d'extraction d'entités nommées comme :

- **l'analyse de dépendance** qui permet d'identifier les dépendances décrivant les relations entre les termes pour déterminer s'ils constituent le sujet, le verbe ou les compléments d'une phrase.
- **la détection des limites des phrases** qui consiste à subdiviser le texte en phrases.

Cependant, avec toutes ces étapes de prétraitement, la reconnaissance d'entités nommées fait face à de nombreuses difficultés à la fois dans le manque de ressources disponibles et dans le domaine de la linguistique.

I.2.2. Quelques problématiques autour de la REN

La reconnaissance d'entités nommées se confronte à de multiples problématiques :

- **La disponibilité des ressources :** le développement de systèmes de reconnaissance d'entités nommées nécessite un ensemble de ressources (corpus annotés, dictionnaires, lexiques, etc.). Cependant, on note surtout une faible accessibilité à ce type de ressources pour certaines langues comme le français. En plus, les données disponibles sont pour la plupart bruitées[13] [4].
- **La désambiguïsation de la majuscule en début de phrase** [1]**:** comme énoncé dans la section I.1.5, la majuscule serait une marque pour la détection d'entités nommées. Mais celle-ci en début de phrase ne garantit pas forcément la présence d'une entité nommée. Donc, d'autres traitements seront nécessaires pour répondre aux interrogations suivantes :

[13] Langage familier, faute d'orthographe, abréviation personnalisée, etc.

- dans quel cas une majuscule en début de phrase peut-elle indiquer la présence d'une entité nommée ?
- quels sont les mots ou termes qui, lorsqu'ils sont en début de phrase, correspondent directement à une entité nommée ?

- **La non exhaustivité :** les entités nommées dépendent fortement du domaine étudié (économique, social, politique, agronomique, géographique, biomédical, etc.). Ce qui fait qu'elles constituent un ensemble non exhaustif.
- **Les problématiques liées à la complexité de la langue :** ces difficultés concernent principalement certains axes de la linguistique notamment l'axe morpho-lexicale et l'axe sémantique [27].
    - **Morpho-lexicale :** cette étape décrit la formation lexicale des mots du texte. On se confronte à la difficulté pour détecter la frontière entre les entités pour les délimiter. C'est surtout le cas pour les entités nommées constituées de plusieurs mots.

    Exemple : « La porte du 3^{iém} millénaire », « Les grilles royales du château de Versailles ».

    - **Sémantique :** elle traite le sens des mots du texte et plusieurs problèmes se posent :
        - **La synonymie** [14]**:** plusieurs expressions peuvent correspondre à une même entité.

        Exemple : « Il habite dans cette maison », « Tu as une nouvelle demeure ».

        Ici, « maison » et « demeure » correspondent à la même entité de type LIEU (LOC en anglais)

        - **L'homonymie et la polysémie :** une même entité peut correspondre à plusieurs types (ou catégories) d'entités.

        Exemple 1 : « La souris vient de sortir de son trou. » (1)
        « La souris ne fonctionne plus. » (2)

        Dans la première phrase, le mot « souris » renvoie à l'animal alors que dans la seconde, elle correspond à un périphérique de l'ordinateur.

Exemple 2 : le mot « **orange** » peut désigner une organisation, un fruit ou bien même une couleur.

- **La métonymie :** une expression peut être associée à une entité qui est différente du type d'entité qu'elle référait habituellement.

Exemple : « Elle lisait un Maupassant ».

Où « Maupassant » signifie « un livre écrit par Maupassant » et non « Maupassant » comme une personne.

I.3. Domaines d'application de la REN

La reconnaissance d'entités nommées (REN) est une tâche utilisée dans plusieurs domaines de la vie quotidienne. Elle est en générale une étape préalable pour d'autres types d'applications du TAL. De plus, elle a un impact significatif sur les performances de ces applications, ce qui la rend cruciale [12]. Son rôle diffère d'une application à une autre.

I.3.1. Indexation et recherche d'informations

L'indexation consiste à représenter un document (texte, vidéo, etc.) par des termes (appelés index) qui renvoient aux sujets dont parle le document [13]**.** Dans une telle application, les entités nommées peuvent servir d'index pour représenter les documents. Par exemple, les mots « clavier, souris, écran, disque dur, etc. » extraits comme des entités nommées, peuvent servir d'index aux documents qui parlent d'ordinateur.

En ce qui concerne la recherche d'informations, elle consiste à trouver les résultats les plus pertinents dans une collection de documents pour une requête donnée. Les entités nommées extraites à partir de la requête permettent de comparer cette dernière aux documents de la collection [14]. Par exemple, dans une requête qui contient le mot « Microsoft » qui correspond à une entité de type Organisation, tous les documents relatifs à Microsoft Corporation seront considérés comme pertinents et seront récupérés.

I.3.2. Systèmes de Questions/Réponses

Un système de questions/réponses (abrégé Q/R) permet de donner une réponse précise et concise à une question émise (requête). Dans de tels systèmes, les entités nommées sont utilisées pour permettre d'identifier le(s) type(s) de réponse(s) attendue(s) [14]. Par exemple, l'entité nommée « orange » peut être classée comme une organisation, un fruit ou bien même

une couleur selon le contexte. Ainsi, la classification appropriée de l'entité nommée permettra de savoir quels sont les documents à cibler pour trouver la réponse adéquate.

I.3.3. Résumé automatique

Il consiste, à partir d'un texte, à donner sa version courte et condensée en gardant le sens et les informations principales. Pour cela, le meilleur moyen est d'utiliser la REN pour extraire et classer les mots pertinents du texte, ce qui permettra de garder le sens du texte.

I.3.4. Fouille d'opinions et analyse de sentiments

La fouille d'opinions consiste à déterminer les caractéristiques décrivant l'orientation de l'opinion d'un texte ou d'une collection de textes. Celle-ci peut être *« objective »* si les caractéristiques relatent *« un fait »* ou bien *« subjective »* si elles relatent *« une opinion »* [5]. Dans ce type d'application, les systèmes de reconnaissance d'entités nommées (REN) serviront d'extracteur de ces caractéristiques sur lesquelles se référaient pour effectuer la fouille.

L'analyse de sentiments s'intéresse à déterminer la polarité d'une opinion. Celle-ci est soit positive, soit négative ou neutre [28]. Ainsi, les éléments pertinents que sont les entités nommées sont les plus utiles pour l'aide à l'analyse et à la prise de décision pour déterminer la polarité des opinions.

I.3.5. Veille technologique, économique, politique

Elle permet, à travers une base de connaissances formée à partir des informations recensées, d'aider à la prise de décision, à la mise en évidence des éléments pertinents dans les textes et à l'estimation du degré de menaces ou d'opportunités d'une activité [1]. Dans ce cas, les entités nommées seront le socle pour la formation de la base de connaissances.

I.3.6. Domaine biomédical

Dans le domaine biomédical, les entités nommées biomédicales participent grandement à l'analyse des données textuelles sur les maladies et sur les génomes, au remplissage de bases de gènes et/ou de maladies, à la détermination des interactions entre les gènes, les liens entre les gènes, etc.

I.3.7. Analyse des courriers pour le support en ligne

Une bonne gestion des courriers dans une entreprise permet d'avoir une meilleure organisation et une notification instantanée et sans échéance. Les entités nommées servent de levier sur lesquelles se basent ces systèmes pour permettre l'orientation des demandes vers le bon service, à améliorer la productivité des employés, à pallier aux problèmes du style de rédaction des courriers électroniques, etc [8].

Par ailleurs, la reconnaissance d'entités nommées (REN) est une tâche bien utile dans tant d'autres domaines du traitement automatique des langues et dans des domaines de la vie active (géographie [22], agronomie, etc.).

Conclusion

Dans ce chapitre, nous avons d'abord examiné l'historique de la REN, puis clarifié les différentes définitions proposées dans la littérature et présenté les formes et types des entités nommées. Ensuite, nous avons exposé les étapes et les problématiques de la REN. Et enfin, nous avons cité quelques domaines d'applications de la REN.

Ces dernières décennies, la reconnaissance d'entités nommées (REN) a fait l'objet de plusieurs convoitises dans le traitement du langage naturel. En raison de son importance dans différents systèmes du TAL, optimiser et évaluer leurs performances reste d'actualité. Nous allons ainsi aborder dans le chapitre suivant les approches utilisées, les métriques d'évaluation et présenter quelques systèmes de REN existants.

Chapitre 2 : Approches, métriques et outils d'extraction d'entités nommées

Introduction

La production de plus en plus croissante de grandes masses de données non structurées a attiré l'attention de la communauté du TAL. L'utilisation des technologies du TAL notamment la reconnaissance d'entités nommées (REN) sur ces données fait l'objet d'une grande attention. La REN nécessite un ensemble de méthodes et de traitements. Ainsi, plusieurs approches ont été proposées dans la littérature et elles peuvent être classées en trois catégories : les approches symboliques ou à base de règles, les approches à base d'apprentissage automatique et les approches hybrides. Chacune de ces approches présente un ensemble d'avantages et de limites. Plusieurs métriques sont utilisées pour mesurer les performances des systèmes de REN.

Dans ce chapitre, nous allons d'abord présenter les différentes approches utilisées pour la REN ainsi que leurs avantages et leurs limites, ensuite décrire les mesures utilisées pour évaluer les systèmes de REN et enfin présenter quelques outils de REN existants.

II.1. Différentes approches de REN

La mise en place d'un système d'extraction d'entités nommées nécessite un ensemble de techniques et de méthodes. Ainsi, différentes approches ont été proposées. Elles peuvent être subdivisées en trois (3) catégories [29] [30]: l'approche à base de règles, l'approche à base d'apprentissage automatique et l'approche hybride.

II.1.1. Approche à base de règles

L'approche à base de règles appelée aussi approche orientée connaissance consiste à mettre en place un ensemble de règles qui permettront de qualifier un mot ou un groupe de mots d'entité nommée éligible dans un domaine ou une langue donnée [20]. Cependant, l'élaboration de ces règles demande une certaine expertise du domaine pour établir les conditions nécessaires qui facilitent le repérage d'entités nommées dans les textes. En plus, elles sont établies à différents niveaux de traitements (morpho-lexicale, syntaxique et sémantique) [10]. Elles sont basées soit sur un ensemble de lexiques qui doivent correspondre aux unités de textes soit sur un dictionnaire renfermant un ensemble de mots. Par ailleurs, un système utilisant les deux, donne de meilleurs résultats [10].

Exemples de règles :

- la présence d'un prénom suivie d'un nom propre qui commence par une majuscule indiquera un nom de personne (Ex : 'Barack Obama') ;
- un mot inconnu + '.Inc', indique une entité de type organisation (Ex : 'Design.Inc') ;
- un terme écrit sous la forme de quatre chiffres indiquera une entité de type date (Ex : '1915').

II.1.1.a. Avantages

L'approche à base de règles présente plusieurs avantages :

- elle a généralement une bonne précision dans la REN [31];
- sa mise en place ne nécessite pas de données annotées.

II.1.1.b. Limites

Néanmoins, l'approche à base de règles présente certaines limites :

- problème de portabilité et de robustesse [4] : un système à base de règles est difficilement généralisable dans un autre domaine autre que celui pour lequel il a été conçu ;
- le travail manuel effectué pour l'élaboration des règles est coûteux en temps de travail ;
- un rappel généralement faible : un système à base de règles peut omettre de potentielles entités si leurs formes ne sont pas couvertes par les règles [32] ;
- elle n'est pas adaptée pour certains textes qui ne respectent pas rigoureusement les règles grammaticales (Exemples : mails, messages, tweets, etc.).

II.1.2. Approche à base d'apprentissage automatique

L'approche à base d'apprentissage automatique (ou *machine learning (ML)* en anglais) est aussi dite orientée données. Elle est entièrement basée sur l'utilisation des algorithmes d'apprentissage et sur un ensemble de données annotées (corpus d'entrainement annotés) [33].

On appelle « corpus », les données textuelles utilisées pour le traitement automatique du langage naturel. Les données textuelles annotées, quant à elles sont appelées « corpus annotés ». Un corpus annoté est un ensemble de documents avec leurs étiquettes d'entités établies selon une typologie donnée. Ces données sont soit fournies par des experts (golden standard corpus[14])

[14] C'est un corpus annotées à la main par plusieurs annotateurs et révisé par un spécialiste de la matière

soit générées automatiquement (silver standard corpus[15]). Celles-ci sont soumises au système pour qu'il apprenne les connaissances et les règles émises pour pouvoir déterminer et bien classer de nouvelles données.

Ces algorithmes sont classés en trois types d'apprentissages [12]:

- **apprentissage supervisé :** il permet à une machine d'apprendre à partir d'un ensemble de données annotées (étiquetées) et de reproduire ces mêmes exemples de sorties. Cependant, les systèmes qui utilisent ce type d'apprentissage prennent la REN comme une tâche de classification.

  Plusieurs algorithmes de ce type sont utilisés dans la REN : *CRF* (Conditional Random Field) [29], *HMM* (Hidden Markov Model), *MEMM* (Modèle de Markov à Entropie Maximale), *SVM (Support Vector Machine)*, etc.
- **apprentissage non-supervisé** : ce type d'apprentissage, ne nécessite pas de données étiquetées [34]; il permet d'extraire des informations sur un ensemble de données dont aucun attribut n'est plus important qu'un autre. Pour ce faire, les données sont séparées ou divisées en plusieurs groupes suivant leurs traits de similarité. Parmi ces algorithmes, on peut citer le plus courant : le **K-means**.
- **apprentissage semi-supervisé** : ce type d'apprentissage, quant à lui, combine les deux premiers (données étiquetées et données non étiquetées). Un système utilisant ce type d'apprentissage doit être capable de :
  - apprendre de nouvelles connaissances sur de nouvelles données ;
  - se passer des données d'origine pour entrainer un nouveau classificateur et de préserver les connaissances déjà acquises.

[15] C'est un corpus généré automatiquement sans l'aide d'annotation

La figure suivante illustre ces différents types d'apprentissages :

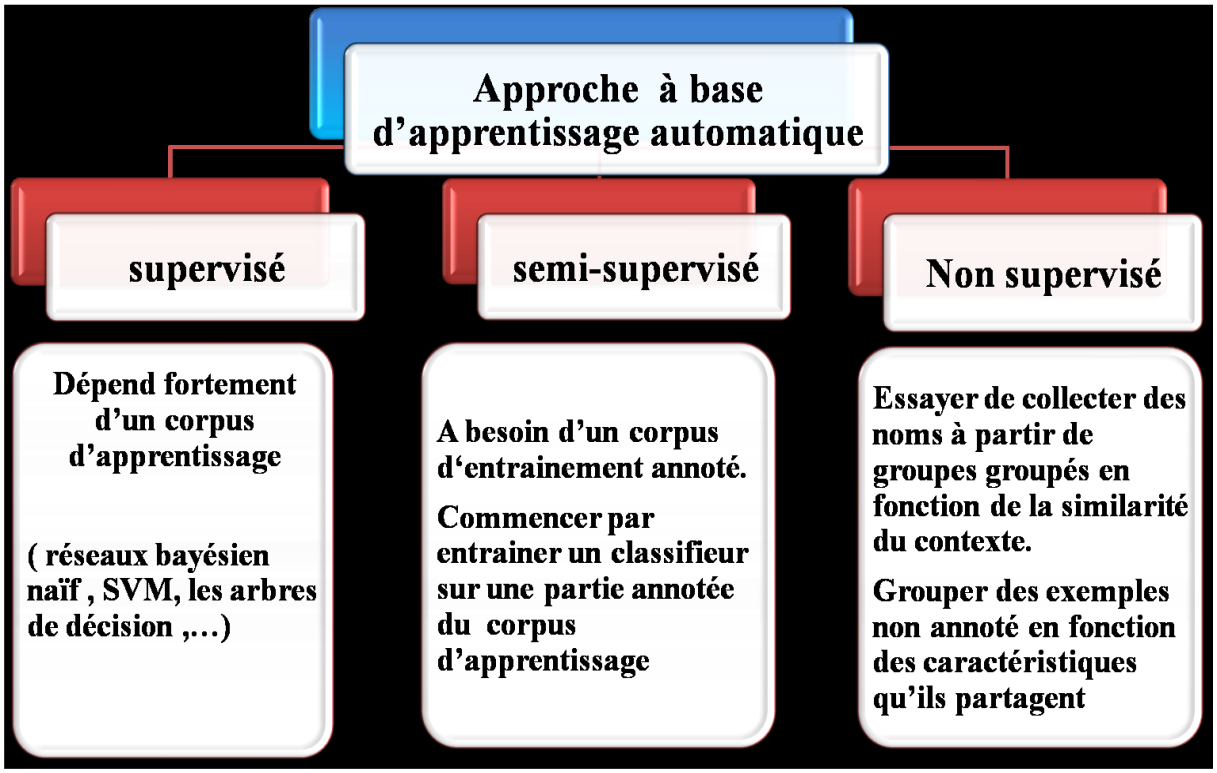

Illustration 6 : Les types d'apprentissages à base automatique [12]

L'approche à base d'apprentissage utilise des classificateurs qui sont exigeants en matière de données d'entrée. La grande puissance de calcul qu'offre la technologie, a donné naissance à une autre forme d'apprentissage automatique moderne appelée apprentissage profond[16] ou apprentissage à base de réseaux de neurones (deep learning, DL en anglais) [4] [35]. Ce dernier, inspiré du fonctionnement du cerveau humain, utilise différentes couches neuronales sur des données brutes pour effectuer ses fonctions prédictives en vue d'apporter plus de fiabilité et de performance [24], ce qui lui confère sa popularité. Il présente différentes architectures notamment [36]: RNN (Recurrent Neural Network), CNN (Convolutional Neural Network), LSTM (Long Short-Term Memory). Les architectures de réseaux de neurones sont construites à partir de deux facteurs : un encodeur pour prendre la représentation du contexte des données en entrée et un décodeur de balise. Ces dernières années sont marquées par de nombreuses recherches sur la combinaison de plusieurs architectures de réseaux de neurones ou bien leur combinaison avec des algorithmes de machine learning. C'est le cas de Chiu qui utilise le LSTM

[16] L'apprentissage profond permet à la machine de construire des concepts complexes sur la base de simples concepts

et le CNN [37], de Huang qui utilise une architecture de réseaux de neurones LSTM-Bidirectionnels comme encodeur et un CRF comme décodeur de balise [38].

On note également l'utilisation des modèles de langues pré-entrainés comme Bert [4] [39].

Par ailleurs, les systèmes utilisant l'approche à base d'apprentissage automatique prennent la tâche d'extraction d'entités nommées comme étant une tâche de classification, en attribuant à chaque token du texte une annotation qui spécifie sa position au sein ou non d'une entité. Ce type d'annotation est spécifique aux entités nommées composées de plusieurs mots. Parmi ces schémas d'annotations [34] [40], le schéma d'annotation **BIO** (Begin, Inside et Outside) est le plus répandu. Ce dernier est utilisé depuis la campagne d'évaluation CoNLL [12] et est considéré comme un standard pour les CRF [40]:

- le premier token d'une entité nommée de type **T**, sera affecté à la classe « B-T » (Begin ou début en français) ;
- pour une entité nommée constituée de plusieurs tokens, on affecte aux tokens suivants le premier token à la classe « I-T » (Inside ou à l'intérieur en français) ;
- et les autres tokens non étiquetés (ne faisant pas partie d'une entité nommée) seront affectés à la classe « O » (Outside ou en dehors en français).

Exemple : « En 2007, Abdoulaye Wade dirigeait le Sénégal. »

Token	Format BIO
En	O
2007	B-DATE
,	O
Abdoulaye	B-PERSON
Wade	I-PERSON
dirigeait	O
le	O
Sénégal	B-LOC
.	O

Tableau 7 : Représentation d'une annotation sous le format BIO

II.1.2.a. Avantages

L'approche à base d'apprentissage automatique présente de nombreux avantages :

- elle permet d'avoir un gain de temps : on passe moins de temps pour détecter et extraire les entités nommées avec un modèle (système) déjà entrainé [29];
- elle a de bonnes performances surtout la précision ;
- elle est facilement adaptable à un autre domaine différent de celui utilisé. Il suffit simplement de fournir de nouvelles données annotées dans le domaine où l'on souhaite faire l'extension ; ce type d'apprentissage est appelé apprentissage par transfert [35] [41].

II.1.2.b. Limites

L'approche à base d'apprentissage automatique se confronte à certaines limites :

- l'annotation des données a un coût ; elle est réalisée par des experts et ces données serviront pour la plupart de données d'entrainement ou de test pour ces modèles [17];
- sa performance sur des données bruitées est limitée ;
- l'aspect « boite noire » des algorithmes d'apprentissage constitue un frein majeur pour une amélioration de leurs performances [32].

II.1.3. Approche hybride

L'approche à base de règles a montré ses performances sur des données canoniques avec une bonne précision par rapport à l'approche à base d'apprentissage automatique [29]. Mais, celle-ci est plus robuste et obtient un meilleur rappel. Ainsi, l'approche hybride combine ces deux pour en tirer le meilleur [33] [42], soit en apprenant au système de façon automatique l'ensemble des règles et connaissances nécessaires puis de faire une revue manuelle sur celle-ci, soit en élaborant les règles à la main et puis corriger et améliorer automatiquement.

II.1.3.a. Avantages

Les grands atouts de cette approche sont puisés dans les deux premières à savoir combiner les points forts de ces approches (établir des règles manuellement et apprendre au système pour qu'il puisse faire la classification de nouvelles données automatiquement).

II.1.3.b. Limites

Les entités nommées sont réputées être une classe indéfinie (ouverte), car elles dépendent entièrement de la langue, du domaine d'application et des types d'entités à extraire. Ceci fait que de nouvelles entités peuvent être éventuellement trouvées selon le type de l'application, ce qui sera toujours défavorable pour la précision d'un système hybride [41].

II.2. Métriques d'évaluation d'un système de REN

Les systèmes d'extraction d'entités nommées sont évalués en utilisant des jeux de données standards pour mesurer leurs performances. Ainsi, plusieurs métriques existent dans la littérature : le Slot Error Rate (SER) [43] utilisé lors de la campagne d'évaluation de Quaero [44] et ETAPE, la Entity Tree Error Rate (ETER) [45]. Mais les métriques les plus répandues en TAL reste la précision (P), le rappel (R) et la F-mesure (F).

- La **précision** permet de déterminer parmi les entités trouvées celles qui sont correctement annotées. Elle est la mesure de la qualité du système [46]. Elle est définie par la formule suivante :

$$Précision = \frac{nombre\ d'entités\ correctement\ étiquetées}{nombre\ d'entités\ étiquetées}$$

Formule 1 : Formule de calcul de la Précision *[17]*

- Le **rappel** est le rapport du nombre d'entités trouvées et correctement étiquetées sur l'ensemble des entités identifiées. Il est appelé aussi la mesure de quantité [46] et est défini par la formule suivante :

$$Rappel = \frac{nombres\ d'entités\ correctement\ etiquettées}{nombres\ d'entités\ à\ trouver}$$

Formule 2 : Formule de calcul du Rappel *[17]*

- La **F-mesure** correspond à la synthèse de la précision et du rappel. Elle est donnée par la formule suivante :

$$F - mesure = \frac{2 * precision * rappel}{precision + rappel}$$

Formule 3 : Formule de calcul de la F-mesure *[17]*

Lors de l'évaluation des systèmes de REN, on note que ces derniers sont confrontés à deux types d'erreurs courantes [15]:

- soit le système donne en sortie moins de résultats que ce qui est attendu ;
- soit plus de résultats que ce qui est attendu.

On parle de **silence**, si le système omet des résultats attendus et de **bruit** si celui-ci en trouve plus que ce qui est attendu.

Le bruit et le silence sont déterminés par les formules suivantes :

$$Bruit = 1 - Precision$$

Formule 4 : Formule de calcul du bruit *[47]*

$$Silence = 1 - Rappel$$

Formule 5 : Formule de calcul du silence *[47]*

Ainsi, dans un système d'extraction d'entités nommées, plus la précision augmente plus le bruit se réduit et plus le rappel augmente plus le silence diminue.

Par exemple, essayer d'améliorer la précision d'un système implique une diminution du bruit (amoindrir le nombre de résultats trouvés). C'est ce qui peut entrainer une omission de certains résultats attendus et donc diminuer le rappel. Ainsi, les résultats d'un système de REN sont bons[17] s'ils sont de qualité (*précision*), suffisamment exhaustifs (*rappel*) et robustes (tolérance au *bruit*) [14].

II.3. Quelques outils d'extraction d'entités nommées

Depuis l'ère des conférences MUC 6 et MUC 7 marquant la naissance de la tâche d'extraction d'entités nommées, plusieurs systèmes[18] de REN ont vu le jour. Ils se différencient par le domaine d'application, par les langues traitées, par les types d'entités qu'ils permettent d'extraire, par leurs performances, par le type d'approches utilisé, etc.

[17] La valeur de la précision et du rappel varie entre 0 et 1 (et pour avoir le résultat en pourcentage, on multiplie leurs valeurs(R et P) par 100

[18] Certains outils font d'autres tâches en dehors de la reconnaissance des entités nommées et pour plusieurs langues aussi, mais nous nous limitons juste à cette étape de traitement et pour la langue française. Il nous est difficile de faire une comparaison entre les outils dont nous parlons ici, car :

- les résultats obtenus par chacun des outils cités ne sont pas donnés de façon homogène (rappel seul ou précision seule, évalué sur plusieurs corpus, etc.).
- les données textuelles utilisées ne sont pas forcément les mêmes.
- certains systèmes n'ont pas été évalués ou en tout cas, on n'a pas trouvé d'article contenant leur évaluation.
- l'évaluation est donnée en fonction des types d'entités extraites.

Mais toute fois leurs résultats sont donnés lorsqu'ils ont été communiqués

Nous allons présenter quelques systèmes de REN de la langue française classés selon l'approche utilisée :

➢ Approche à base de règles

- **CasEN** [31]: c'est un système de REN par cascades de transducteurs implémenté avec le système CaSys de Friburger et al. [31] et disponible sur la plateforme Unitex[19] [48]. Ce système a été proposé lors de la campagne d'évaluation Ester 2 et s'appuie sur des dictionnaires à large couverture dont Prolex. Cet outil a été évalué sur le corpus Eslo 1 (précision 92 % et rappel 88,4 %) et lors de la campagne d'évaluation Ester 2 (précision 79,3 % et rappel 65,8 %).
- **Nemesis** [49]: c'est un système de REN pour le français. Il est basé sur une approche à base de règles s'appuyant sur les indices internes définis par McDonald et sur des patrons basés sur les étiquettes sémantiques correspondants aux lexiques. Il a été évalué sur un corpus composé de textes extraits du journal français *Le Monde* et de pages Web, où il obtient une précision de 95% et un rappel de 90% pour la reconnaissance des patronymes et des organisations.

➢ Approche à base d'apprentissage automatique

- **SEM**[20] [3]**:** c'est un outil qui permet d'effectuer plusieurs tâches notamment le Part-Of-Speech, le chunking et l'extraction d'entités nommées. Cet outil est basé sur l'utilisation des CRF (Conditional Random Field) pour effectuer l'annotation avec l'outil Wapiti[21] et est entrainé sur le corpus French Treebank[22]. Son utilisation sous forme de pipeline lui permet de traiter des textes bruités. Le module d'extraction d'entités nommées de SEM obtient lors de son évaluation par un processus de validation croisée à 5 plis une précision de 86,38%, un rappel de 80,30% et un F-mesure de 83,23%.

[19] https : //www-igm.univ-mlv.fr/˜unitex/
[20] Il est fourni avec une interface web disponible à l'adresse suivante : http://apps.lattice.cnrs.fr/sem/
[21] https://wapiti.limsi.fr/manual.html
[22] http://ftb.linguist.univ-paris-diderot.fr/

- **LiaNE** [32]: c'est un système de REN développé dans le cadre de la campagne d'évaluation ESTER. Il est basé sur une approche doublement faite à base d'apprentissage automatique :
    - avec un processus génératif à base de HMM pour prédire les étiquettes syntaxiques (POS) et sémantiques des mots du texte,
    - et un processus discriminant à base de CRF pour déterminer les bornes des entités et leurs catégories en utilisant le modèle BIO présenté en II.1.2. Ce système a obtenu sur le corpus ESTER des précisions et des rappels de 81,6 et 80,7 ; 51,4 et 55,5 ; 81,3 et 88,4 respectivement pour les entités de types LOC, ORG et PERS.
- **Stanford CoreNLP**[23]**:** c'est un outil implémenté en java pour le traitement de textes (tokenisation, étiquetage, REN, etc.). Il peut être utilisé en ligne de commande via son API de programmation ou via un serveur web. Il est disponible en anglais comme dans d'autres langues telles que le français, l'allemand, l'arabe, l'espagnol et le chinois. Stanford CoreNLP est aussi disponible dans d'autres langages de programmation (Python, PHP, JavaScript, etc.), il suffit juste d'utiliser l'API correspondante pour la configuration du serveur.
- **Spacy**[24]**:** c'est une bibliothèque python open source pour le traitement automatique des langues. Il fournit un module de tokenisation, un étiqueteur morpho-syntaxique (POS), un module de REN, etc. Il dispose de modèles entrainés dans plusieurs langues (français, anglais, allemand, espagnol, portugais, etc.). Spacy est basé sur une approche d'apprentissage statistique utilisant les CNN (réseaux de neurones convolutifs) pour reconnaitre les entités nommées. Il permet aussi d'élaborer des modèles à base de règles en créant des motifs basés sur les tokens avec l'importation de l'outil « Matcher ».
- **SPARK NLP**[25]**:** c'est une bibliothèque open source destinée au traitement avancé de textes (TAL) en utilisant les langages python, scala, ou java. Il supporte le traitement de plusieurs langues notamment l'anglais et le français. Il

[23] https://stanfordnlp.github.io/CoreNLP/
[24] https://spacy.io/
[25] https://nlp.johnsnowlabs.com/

fournit des canaux de traitement (tokenizer, stemmeur, lemmatiseur, NER, etc.) avec des modèles de réseaux de neurones pré-entrainés. SPARK NLP présente une extension commerciale (SPARKL NLP for Healthcare) pour exploiter les textes cliniques et biomédicaux (reconnaissance d'entité nommée clinique, liaison entre entités, etc.). Aussi, il possède une extension permettant la reconnaissance optique de caractères (SPARK ORC) à partir d'images, de documents numériques. Il permet de faire entre autre l'extraction de texte à partir d'image, le recadrage d'images, la suppression d'objet en arrière-plan, etc. Ainsi, Spark NLP se présente comme étant un outil de traitement automatique multi-domaine (domaine général, biomédical et OCR) et un outil multilingue du TAL.

- **Open NLP[26]:** c'est une boite à outils open source java pour le traitement automatique des langues. Il supporte des tâches telles que la tokenisation, la segmentation de textes en phrases, l'extraction d'entités nommées, etc. Il est basé sur une approche d'apprentissage automatique basée sur la machine de Markov à une entropie maximale (MEMM). Ces fonctionnalités sont accessibles via son API et il est multilingue.

➢ **Approche hybride :**

- **mXS** [32]: ce système d'extraction d'entités nommées a été proposé lors de la campagne d'évaluation ETAPE. Il utilise une approche hybride (approche à base de données combinée à une approche orientée connaissance reposant sur la recherche à partir des données d'apprentissage et des motifs). Il a obtenu lors de la campagne d'évaluation ETAPE une précision de 76,4% et un rappel de 62,3%.

[26] https://opennlp.apache.org/

Conclusion

Dans ce chapitre, nous avons présenté dans un premier temps les différentes approches utilisées pour la REN. Nous avons ensuite décrit des métriques d'évaluation permettant de mesurer les performances de tels systèmes. Enfin, des outils de REN sont présentés. Ces méthodes et ces approches, sont-elles applicables dans les domaines de spécialités notamment le domaine biomédical ? Existe-t-il des outils spécifiques au domaine biomédical ? Quelles performances de ces outils pour le traitement des données biomédicales ? Les réponses à ces questions feront l'objet du prochain chapitre.

Chapitre 3 : Extraction d'entités nommées dans le domaine biomédical

Introduction

L'expansion massive des données (textes, images, sons, vidéos, etc.) disponibles sur internet se reflète aussi sur les données des autres domaines tels que le domaine biomédical où le nombre de documents disponibles connait une fulgurante hausse. Par exemple, la base de données MEDLINE[27] contient plusieurs millions de références d'articles dans le domaine biomédical.

L'exploitation de ces informations devient un besoin majeur et la création de systèmes automatiques pour traiter ces informations est un grand défi. Ces données, qui sont produites par des spécialistes du domaine biomédical, sont pour la plupart du temps non structurées. Une telle problématique ne facilite pas leurs exploitations. Cependant, la reconnaissance d'entités nommées est une tâche importante pour beaucoup d'autres applications du domaine biomédical telles que l'aide à la prise de décisions médicales, la recherche d'informations médicales [40] [55] [56], l'extraction des relations entre les entités médicales [2] [35] [57], etc.

Plusieurs travaux se sont intéressés à l'exploitation de ces ressources, en identifiant les informations relatives aux entités biomédicales (noms de gènes, de maladies, de médicaments, etc.) et leurs relations. La fouille de textes dans ce domaine est appelée le traitement du langage biomédical (BioNLP en anglais), tandis que la reconnaissance d'entités nommées est dite reconnaissance d'entités nommées biomédicales (BioNER en anglais).

Ce chapitre est structuré en trois sections. D'abord, la première section s'intéresse à la tâche même de l'extraction d'entités nommées biomédicales en définissant cette tâche, en citant quelques types d'entités biomédicales et les types d'encodage utilisés. Ensuite, dans la deuxième section, nous exposons les difficultés rencontrées et les défis à relever dans ce domaine notamment pour la langue française. Enfin, dans la troisième section, quelques outils existants et des données disponibles dans le domaine biomédical sont décrits.

[27] C'est une base de données internationale pour les sciences de la santé et les sciences biomédicales

III.1. Reconnaissance d'entités biomédicales

III.1.1. Définition

Une entité nommée biomédicale, appelée aussi entité biomédicale, est définie comme une entité nommée qui se réfère spécifiquement à une instance du domaine biomédical [34]. Ces types d'entités couvrent les noms de maladies, les noms de gènes, de protéines, de médicaments, de dosages, etc. Ces derniers se basent sur les concepts de l'UMLS[28] (Unified Medical Language System, système de langage médical unifié en français).

Le processus d'identification et de catégorisation de ces entités est appelé la reconnaissance d'entités biomédicales (BioNER) [58].

III.1.2. Typologies d'entités biomédicales

Les entités biomédicales concernent toutes les unités de textes qui font référence à une entité du domaine biomédical [47]: noms de gênes, de maladie, d'ADN, de résultats d'examens, de traitements, d'handicaps, etc. Ces types d'entités dépendent du type des données biomédicales à traiter ; celles-ci peuvent concerner les maladies, les protéines, les aliments, etc.

Le tableau ci-dessous illustre quelques exemples d'entités biomédicales et leurs types.

Type d'entité	**Exemple d'entité médicale**
Maladie	Traumatisme lombaire, Alzheimer
Symptôme	Douleur lombaire
Examen	Radiologie
Traitement	Corset
Médicament	Paracétamol, Solupred

Tableau 8 : Exemples d'entités médicales et leurs types

[28] https://www.nlm.nih.gov/research/umls/index.html

III.1.3. Typologies d'encodages d'entités biomédicales

La reconnaissance d'entités nommées est reconnue comme une tâche typique d'étiquetage de séquences dont le seul but est, pour une séquence textuelle donnée, de trouver la séquence d'étiquette correspondante [59]. Ainsi, pour résoudre le problème de segmentation et d'étiquetage des entités, plusieurs types d'encodage ont été proposés. Ces encodages servent à ajouter des étiquettes afin d'identifier si un mot est en début, à l'intérieur ou à l'extérieur d'une entité biomédicale. En effet, ces encodages sont utilisés par les systèmes utilisant une approche à base d'apprentissage automatique.

III.1.3.a. Encodage IO

L'encodage IO est l'un des types d'encodage utilisé en TAL notamment le plus simple d'usage appelé aussi encodage *entrée/sortie* (E/S) [30] [60]. N'ayant que 2 classes (I/O), il utilise la règle suivante :

- une entité nommée simple ou composée (plusieurs tokens) de type T, est affectée à chacun de ces tokens à la classe « I-T » (Inside, intérieur en français) ;
- et un token ne faisant pas partie d'une entité nommée est affecté à la classe « O ».

Le tableau suivant illustre un exemple d'annotation utilisant l'encodage IO.

Tokens	**encodages**
l	O
'	O
hypotension	I
posturale	I
induite	O
dans	O
la	O
maladie	I
de	I
pakinson	I
:	O
une	O
etude	O
longitudinale	O
sur	O
les	O
effets	O
du	O
sevrage	O
medicamenteux	O

Tableau 9 : Exemple d'annotation utilisant le format I/O

Ce type d'encodage est limité dès l'instant où 2 entités nommées se suivent, dans ce cas celui-ci ne pourra pas spécifier ni le début, ni la fin de chacune des entités nommées.

III.1.3.b. Encodage BIO

L'encodage BIO, utilisé depuis la campagne CoNLL, reste le plus répandu et standardisé pour les CRF [40]. Il permet de spécifier le début de chaque entité nommée :

- le premier token d'une entité nommée de type **T** est associé à la classe « B-T » (Begin ou début en français) ;
- pour une entité constituée de plusieurs tokens, chacun de ses tokens suivant le premier est associé à la classe « I-T » (Inside ou à l'intérieur en français) ;
- et les autres tokens non étiquetés (ne faisant partie d'une entité nommée) sont associés à la classe « O » (Outside ou en dehors en français).

Le tableau suivant illustre un exemple d'encodage BIO.

Tokens	**encodages**
l	O
'	O
hypotension	B
posturale	I
induite	O
dans	O
la	O
maladie	B
de	I
pakinson	I
:	O
une	O
etude	O
longitudinale	O
sur	O
les	O
effets	O
du	O
sevrage	O
medicamenteux	O

Tableau 10 : Exemple d'annotation utilisant le format BIO

La principale limite de ce type d'encodage se trouve dans le fait qu'il ne permet pas de montrer la fin de chaque entité nommée.

III.1.3.c. Encodage BMEWO

L'encodage BMEWO, quant à lui, tient compte des tokens frontières des entités nommées (début et fin) [60]. Ainsi, pour une entité nommée de type T, constituée de plusieurs tokens :

- le premier token est associé à la classe « B-T » ;
- les autres tokens suivant le premier token et n'étant pas le dernier token de l'entité seront affectés à la classe « M-T » ;
- le dernier token d'une entité nommée est affecté à la classe « E-T » ;
- pour une entité nommée constituée d'un seul token, celui-ci est associé à la classe « W-T» ;
- et les autres tokens qui ne font pas partie d'une entité nommée sont associés à la classe «O ».

Le tableau suivant illustre un exemple de l'encodage BMEWO.

Tokens	**encodages**
l	O
'	O
hypotension	B
posturale	E
induite	O
dans	O
la	O
maladie	B
de	M
pakinson	E
:	O
une	O
etude	O
longitudinale	O
sur	O
les	O
effets	O
du	O
sevrage	O
medicamenteux	O

Tableau 11 : Exemple d'annotation utilisant le format BMEWO

III.1.3.d. Encodage BILUO

L'encodage BILUO est considéré comme le plus facile à apprendre par rapport à celui de BIO, car il marque explicitement les jetons de frontières [61] [62]. Pour une entité nommée de type **T :**

- son premier token est associé à la classe « B-T » (Begin, début en français) ;
- pour une entité nommée constituée de plusieurs tokens, chacun de ses tokens suivant le premier et diffèrent du dernier token est associé à la classe « I-T » (Inside, à l'intérieur en français) ;
- le dernier token de l'entité est associé à la classe « L-T » (Last, dernier ou fin en français) ;
- pour une entité constituée d'un seul token, ce dernier est associé à la classe « U-T » (Unit, unité en français) ;
- et les autres tokens non étiquetés (ne faisant pas partie d'une entité) sont associés à la classe « O » (Outside ou en dehors en français).

Le tableau suivant illustre un exemple de l'encodage BILUO.

Tokens	**encodages**
l	O
'	O
hypotension	B
posturale	I
induite	O
dans	O
la	O
maladie	B
de	I
pakinson	L
:	O
une	O
etude	O
longitudinale	O
sur	O
les	O
effets	O
du	O
sevrage	O
medicamenteux	O

Tableau 12 : Exemple d'annotation utilisant le format BILUO

III.2. Problématiques et défis de la BioNER

III.2.1. Problématiques de la BioNER

La reconnaissance d'entités biomédicales tout comme la reconnaissance d'entités nommées dans le domaine général est confrontée à de nombreuses difficultés notamment la disponibilité et l'accessibilité des ressources surtout pour la langue française. Face à ces problèmes, on voit de plus en plus d'organismes et de chercheurs qui, soit par le biais d'organisation de campagnes d'évaluation comme le Défi Fouille de Textes (DEFT éditions 2019 et 2020) [55], soit par des recherches [63], contribuent à la création de ces types de ressources. De plus, ces données contiennent souvent des informations personnelles des patients [64] [65]. Avec les lois sur la protection des données à caractère personnel, un ensemble de traitements est obligatoirement effectué sur ces données pour supprimer ou pour anonymiser ces informations personnelles [62] avant une éventuelle publication ou utilisation dans le monde de la recherche. Dans ce cadre, des règles d'anonymisation des dossiers patients sont bien détaillées dans la thèse de Grouin [62]. Par ailleurs, l'accessibilité aux données pour la création de corpus spécialisé se heurte aussi à la difficulté d'obtenir certaines données. Par exemple, il n'est pas facile d'accéder aux dossiers des comptes rendus cliniques [62]. En plus de ces difficultés, on note d'autres problèmes liés au langage médical:

- **Noms d'entités longs :** certaines entités biomédicales sont présentées sous de longues chaines de caractères ; Exemple : *« cellules épithéliales thymiques normales »*
- **Imbrication des entités médicales :** des entités médicales peuvent être imbriquées dans d'autres.

  Exemple : « un patient atteint d'une hypertrophie de la prostate »

  « **hypertrophie** » est une entité médicale de type *signe*

  « **prostate** », de type *organe* ou de type *sexe* (pouvant déterminer le sexe du patient),

  Les deux entités combinées donnent une entité de type *pathologie* ou *maladie* « **hypertrophie de la prostate** ».
- **Noms en commun :** deux ou plusieurs entités peuvent partager un même nom principal.

  Exemple : *« protéines 90 et 73 kda », se réfère à deux protéines qui sont réunies en une seule entité « protéine 90 kda » et « protéine 73 kda ».*
- **Forme orthographique :** le nom d'une entité peut se présenter sous plusieurs formes orthographiques suivant la manière dont elle se présente sur un texte donné.

  Exemple *: « N-acétylcystéine », « N-acétyl-cystéine », « NAcetylcysteine ».*

- **Abréviation ambiguë :** les entités peuvent présenter une abréviation qui peut être interprétée différemment selon le contexte médical.

Exemple *: « TFC » peut designer « facteur de cellule T » ou bien « fluide de culture tissulaire »*.

De plus, les entités biomédicales tout comme les entités nommées sont réputées être des éléments qui participent à la meilleure compréhension du texte, ce qui pourrait nous emmené à se demander lors de leur extraction [66]:

–**d'inclure ou non les articles** (Exemple : *le coronavirus* ou *coronavirus*) et **les possessifs** (Exemple : *sa thérapie cancéreuse* ou *thérapie cancéreuse*)
– **d'inclure ou non les adverbes** (Exemple : *tous les médicaments*) ;
– **d'inclure ou non les pourcentages, les chiffres** (*Exemple : 30 cancers*) et **les doses des médicaments** (*Exemple : amoxyline 1g*) ;
– **d'annoter les abréviations qui suivent les entités médicales** *(Exemple : Virus de l'immunodéficience humaine (VIH)*) les deux ensemble ou chacune à part.

Avec toutes ces difficultés, la tâche de reconnaissance d'entités biomédicales se doit de relever plusieurs défis.

III.2.2. Défis autour de la BioNER

La reconnaissance d'entités nommées dans le domaine biomédical, d'autant plus qu'elle est préalable à la création de nombreuses autres applications aidant à la prise de décision médicale, se doit de relever de plus en plus de défis, notamment :

- **l'extraction d'entités rares et émergentes** [4]: l'un des défis les plus importants de ce domaine à relever reste la reconnaissance d'entités médicales dites *« rares »* et *« émergentes »*. Les *« entités rares »* sont celles dont leurs formes d'écriture sont uniques ; ce type d'entité est appelé en anglais *« singleton entity »*. Les *« entités émergentes »* sont toutes les entités nommées qui ne sont quasiment jamais vues auparavant. Ces deux types d'entités sont appelées les « entités hors vocabulaires ».
- **l'extraction des entités imbriquées** [55]: extraire les entités médicales imbriquées les unes dans les autres se présente comme un des défis majeurs à relever. **Exemple :** « Un patient atteint d'une hypertrophie de la prostate ». On note dans cet exemple la présence de l'entité de type *signe* « **hypertrophie** » et celui de type *sexe* (qui peut permettre de

déterminer le sexe du patient) « **prostate** », et une entité de type *maladie* qui combine les deux premières entités pour former une entité de type *pathologie* « **hypertrophie de la prostate** ». Alors, un système qui permettra d'identifier et de catégoriser toutes ces entités imbriquées relèverait bien ce défi.

- **le manque ou l'insuffisance des données (corpus annoté) :** dans le domaine médical surtout pour la langue française, la création de ressources, qui sont primordiales pour l'entrainement [63] [67] et l'évaluation des modèles, ainsi que leurs anonymisations [62] [68] [69] restent des défis majeurs.
- **l'homogénéisation des données :** les données disponibles dans le domaine biomédical présentent une grande variabilité, du fait qu'elles proviennent de différentes spécialités, de différentes sources, etc. [70] .C'est pour cette raison qu'il est difficile d'envisager un corpus annoté englobant tous les types de textes et tâches du domaine. Alors, un travail de recherche mené en ce sens présentera un grand atout pour faire avancer la recherche sur la reconnaissance des entités biomédicales.

Malgré tous ces défis mentionnés ci-avant, les recherches portant sur la reconnaissance d'entités biomédicales ont permis de développer de nombreux outils, mais également de constituer des corpus de données annotés.

III.3. Outils et jeux de données disponibles

III.3.1. Outils existants

De nombreux travaux de recherches émis dans ce domaine ont donné naissance à plusieurs outils et systèmes d'extraction d'entités biomédicales, mais peu d'entre eux supportent la langue française. Ces outils diffèrent par les types d'entités extraites, par l'approche utilisée (à base de règles, par apprentissage ou hybride), par les corpus utilisés (articles scientifiques, rapports cliniques...), etc.

Dans cette section, nous présentons quelques outils de reconnaissance d'entités biomédicales.

- **SPARK NLP :** c'est une bibliothèque open source destinée au traitement avancé de textes (TAL) supportant les langages de programmation python, scala, et java. C'est un outil multilingue (français, anglais, etc.) et multi-domaines (général, biomédical et OCR). Il utilise des modèles de réseaux de neurones pré-entrainés et fournit son traitement grâce à des canaux de traitement constitué de tokenizers, stemmeurs, lemmatiseurs, ner, etc.

- **cTAKES** [71]**:** cTAKES (Clinical Text Analysis and Knowledge Extraction System) est un outil open source en java qui permet de faire du traitement du langage naturel pour extraire des informations sur des dossiers cliniques (médicaments, signes, symptômes, etc.). Il se base sur le système de langage médical unifié (UMLS) et est créé à l'aide de l'architecture pour l'analyse de contenu non structuré UIMA[29] (Unstructured Information Management Architecture) et de la boite à outil OpenNLP [71].

  Ses composants s'exécutent sous forme d'un pipeline séquentiel allant de la tokenisation à la REN en passant par d'autres étapes telles que l'étiquetage morpho-syntaxique, etc. Il accepte les fichiers sous format XML, du texte bruité et du texte sous format de document clinique. cTakes combine les deux approches (à base de règle et par apprentissage automatique) pour l'extraction d'entités médicales sur les textes cliniques.

III.3.2. Jeux de données

Dans cette section, nous présenterons quelques jeux de données (corpus) biomédicales qui sont disponibles et libres pour la langue française.

- **Le corpus de français médical du CRTT**[30] **(CRTT-MED)** [72]**:** il est constitué d'articles de recherche dans le domaine biomédical. Ces articles sont extraits des revues de la base de données Science Direct, disponible sous format bruité et sous format étiqueté en partie de discours et lemmatisé.
- **Le corpus médical QUAERO**[31] **:** c'est une ressource développée dans le cadre de la recherche pour la reconnaissance d'entités et la normalisation [63]. Ce corpus est composé de titres d'articles extraits de MEDLINE[32] et de documents EMEA (European Medication Agency) [72], annoté manuellement en utilisant les concepts de l'UMLS (Unified Medical Language System) [73].
- **Le corpus clinique MERLoT** [74]: c'est une ressource développée dans le cadre du projet ANR CA-BeRneT comme support de recherche en traitement automatique de la langue clinique en français.

[29] https://uima.apache.org/
[30] http://perso.univ-lyon2.fr/~maniezf/Corpus/Corpus_medical_FR_CRTT.htm
[31] https://quaerofrenchmed.limsi.fr/
[32] https://bib.umontreal.ca/guides/bd/medline

- **Le corpus médical parallèle Français/Anglais EDP**[33]: ce corpus en accès libre est constitué d'articles rédigés en français accompagnés de titres en anglais.
- **Le corpus CAS** [76]**:** c'est un corpus contenant des cas cliniques de patients réels ou fictifs. Il couvre plusieurs spécialités médicales en se focalisant sur différentes situations cliniques. Une partie de ce corpus a été mise à la disposition des participants lors de la campagne d'évaluation DEFT 2019.

Conclusion

Dans ce chapitre, nous avons fait une revue de la littérature sur la reconnaissance d'entités nommées dans le domaine biomédical. D'abord, nous avons défini une entité nommée biomédicale, puis cité quelques types d'entités biomédicales et donné les types d'encodages utilisés. Ensuite, les problématiques et les défis de la BioNER sont décrits. Et enfin, nous avons cité des outils du domaine et des données existantes.

Dans le quatrième chapitre, nous allons présenter notre méthodologie adoptée pour extraire les entités nommées sur des données biomédicales, décrire les expérimentations réalisées et donner les résultats obtenus.

[33] https://cabernet.limsi.fr/EDP.html

Chapitre 4 : Proposition d'une méthode d'extraction d'entités nommées

Introduction

La reconnaissance (extraction) d'entités nommées (REN) particulièrement dans le domaine biomédical fait depuis des années l'objet de beaucoup de recherches. Elle est utilisée dans de nombreuses applications aidant à la prise de décision médicale. Bien que pour les langues telles que l'anglais les ressources (corpus annotés) et outils de REN sont bien fournis, ils doivent être améliorés pour d'autres langues comme le français. Le développement d'un système de reconnaissance d'entités nommées (REN) requiert plusieurs ressources suivant l'approche adoptée : les types d'entités de sortie, un corpus annoté pour l'approche supervisée (à base d'apprentissage automatique), des règles, des lexiques et des dictionnaires pour l'approche symbolique [4] [77] [78]. Toutefois, chacune de ces ressources joue un rôle particulier :

- le corpus annoté sert de données d'entrainement ou d'évaluation ;
- les types de sortie servent à déterminer la nature sémantique des entités ;
- et les lexiques et dictionnaires, eux permettent de fournir les informations sur les entités à extraire.

Dans ce chapitre, nous allons d'abord présenter la méthodologie adoptée pour l'extraction d'entités nommées de types *« âge »* et *« genre »* sur des données biomédicales (cas cliniques) en français. Ensuite, les corpus, les métriques et les outils utilisés pour évaluer notre approche seront détaillés avant de décrire et d'analyser les résultats de cette dernière.

IV.1. Méthodologie

Notre approche permet d'extraire les entités nommées de types *« âge »* et *« genre »* à partir de données cliniques. Il s'agit d'une approche à base de règles qui utilise des lexiques, des dictionnaires de mots et des règles écrites manuellement. Ces règles s'accompagnent d'un traitement à l'aide de mots déclencheurs qui permettront d'affiner les résultats. Nous proposons un ensemble de règles et utilisons respectivement des dictionnaires pour extraire l'ensemble des entités de type *« âge »* et des lexiques pour extraire les entités de type *« genre »* à partir de données cliniques.

IV.1.1. Les règles et les dictionnaires

Les règles et les dictionnaires sont utilisés pour extraire l'âge du (des) patient(s) dans des données cliniques, mais il y a des cas où l'âge du patient n'est pas spécifié. Pour cela, un ensemble de règles ont été définies comme suites :

- si l'âge est indiqué en années, nous récupérons seulement l'âge associé à l'individu. Par exemple, avec « un patient de 26 ans », nous récupérons l'âge qui correspond à *« 26 »* ;
- si l'âge est indiqué en mois, nous convertissons le nombre de mois en années. Par exemple, avec « un enfant de 16 mois », nous arrondissons l'âge à 1 an ;
- si l'âge est écrit en lettres (en mois ou en années), nous utilisons un dictionnaire qui permet de l'associer à son écriture chiffrée correspondante. Par exemple, le dictionnaire suivant ['dix', 'vingt', 'trente',…, 'cent'] est associé au dictionnaire chiffré suivant ['10', '20', '30',…,'100'] ;
- si un adjectif, comme « quadragénaire » est donné pour indiquer l'âge du patient, un dictionnaire est également utilisé pour parer à cette éventualité. Par exemple, ['vingtenaire', 'trentenaire',…,'centenaire'] est associé au dictionnaire suivant ['20','30', …,'100'];
- si plusieurs patients sont présents dans le cas clinique, on fait une extraction des déclencheurs comme « âgées de », « âges respectifs », « âgés respectivement », etc.;
- dans le cas où l'âge du patient n'est pas spécifié, la valeur « NUL » est associée à ce cas.

IV.1.2. Les lexiques

Les lexiques sont utilisés pour extraire le genre du patient décrit dans le cas clinique donné. Le genre d'un patient dans un cas clinique est soit *masculin* pour spécifier un patient de sexe masculin ou *féminin* pour un patient de sexe féminin. Par ailleurs, le genre *masculin/féminin* correspond à un cas qui décrit plusieurs patients de sexes différents.

Deux lexiques ont été utilisés pour extraire les informations relatives aux genres :

- le premier lexique est relatif aux mots qui désignent une personne de genre « féminin » : *« sexe féminin », « femme », « madame », « fille », « fillette », etc.* Nous avons également ajouté des mots déclencheurs liés à l'accord des verbes employés *« née », « âgée », etc. ainsi* que des déclencheurs liés à une partie du corps comme *« vagin », « ovule », etc.*

- le deuxième lexique est relatif aux mots qui désignent une personne de genre « masculin » *: « sexe masculin », « patient », « garçon », « homme », etc*. Nous avons également ajouté des mots déclencheurs liés à l'accord des verbes employés *« né », « âgé », etc.* et des déclencheurs liés à une partie du corps comme *« testicule », « prostate », etc.*

Ces lexiques sont utilisés pour déterminer le nombre d'occurrences correspondant à chaque genre. Ainsi, si les mots trouvés appartiennent tous au lexique « masculin », ce genre sera associé directement au cas clinique, de même pour le genre « féminin ». En revanche, si nous avons des mots qui appartiennent aux deux lexiques, le genre majoritaire l'emporte et il sera associé au cas. En outre, le genre « masculin/féminin » est obtenu que s'il y a autant de nombre de mots appartenant aux deux lexiques.

La figure ci-dessous présente l'architecture de notre approche de REN :

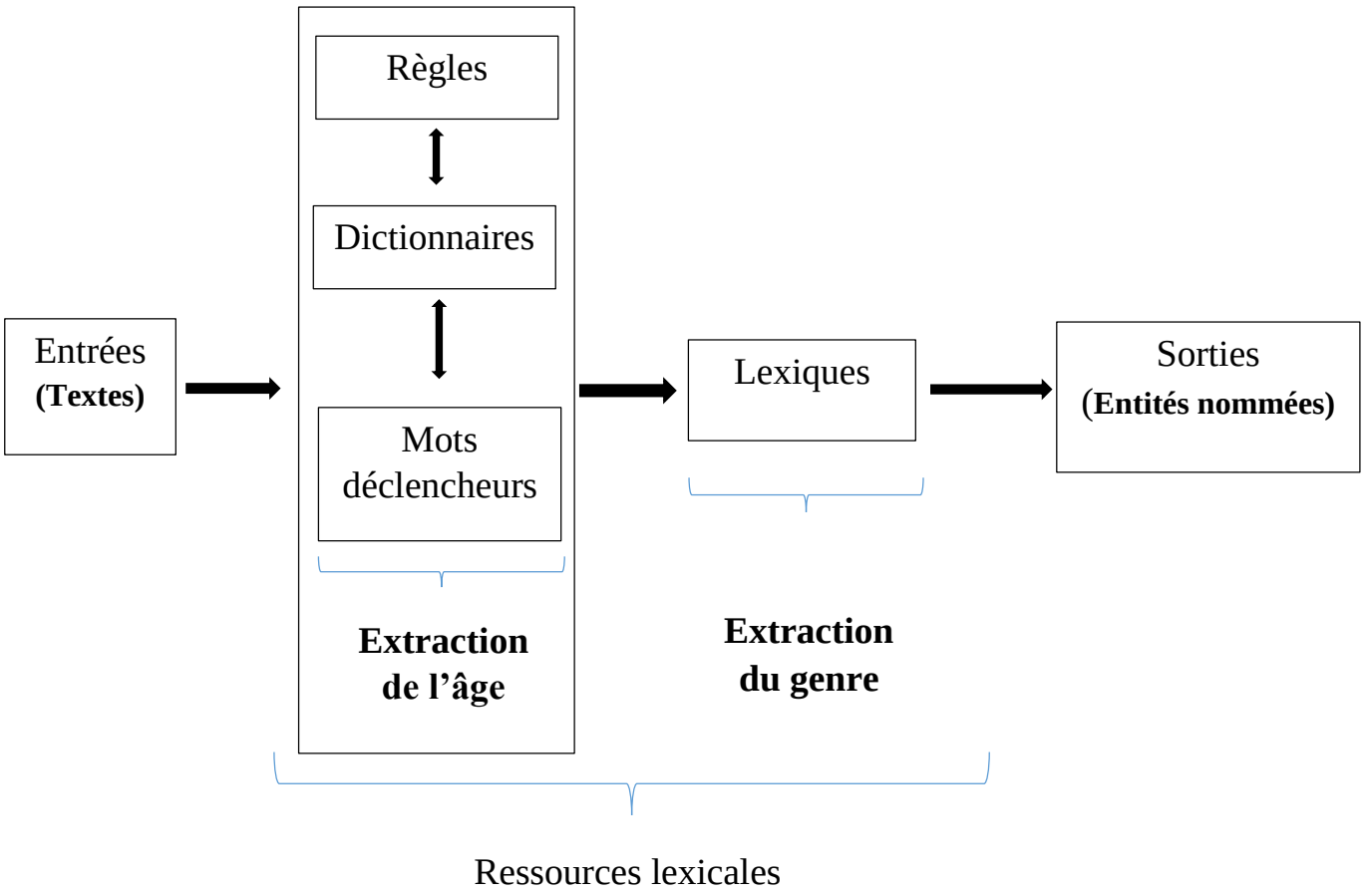

Illustration 7 : Architecture de notre approche de REN

IV.2. Expérimentations

IV.2.1. Corpus

Les données textuelles utilisées pour le traitement du langage naturel sont appelées « corpus » et les données textuelles annotées, quant à elles sont appelées « corpus annotés ». Un corpus annoté est un ensemble de documents avec leurs étiquettes d'entités établies selon une typologie donnée. Nos travaux ont été réalisés sur les corpus fournis lors de la campagne d'évaluation DEFT 2019[34] (Défi de Fouille de Texte édition 2019) [76] [79] [80] [81] [82]. Cette campagne d'évaluation porte sur l'analyse de cas cliniques rédigés en français. Les types d'entités nommées à annoter sont regroupés en quatre catégories principales : âge, genre (masculin, féminin), issue et origine de l'admission. Ces cas cliniques décrivent la situation des patients qui sont soit réels soit fictifs ; ils sont issus de différentes sources (scientifique, juridique, didactique, associative, etc.). Le jeu de données d'entrainement fourni par les organisateurs est constitué de 290 cas, tandis que le jeu de test est composé de 427 cas.

Le genre a été réparti comme suit dans le corpus d'entrainement :

Féminin	Masculin	Féminin/masculin	N/A
117	168	4	1

Tableau 13 : Répartition du genre dans le corpus *[79]*

La figure 13 montre un exemple d'annotation de la campagne DEFT 2019.

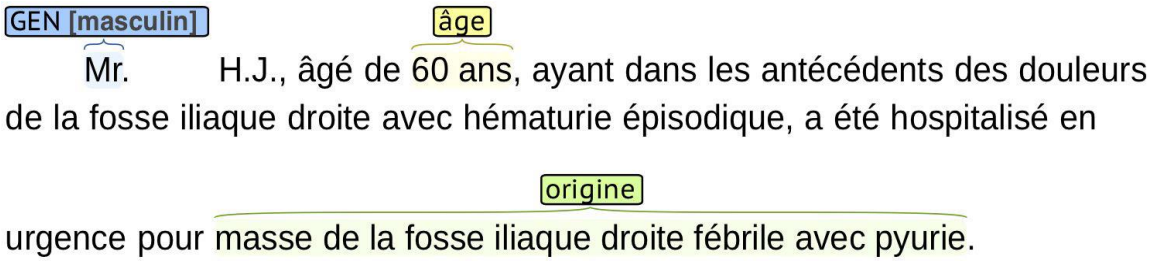

Illustration 8 : Exemple d'annotation de la campagne DEFT 2019 [79]

IV.2.2. Métriques d'évaluation

Plusieurs mesures sont proposées pour évaluer les systèmes d'extraction d'entités nommées ; elles permettent d'évaluer la capacité d'un système à extraire les entités nommées attendues.

[34] https://deft.limsi.fr/2019/

Nous avons utilisé les mêmes métriques utilisées lors de la campagne d'évaluation DEFT 2019 ; la précision, le rappel et la F-mesure (cf II.2).

IV.2.3. Technologies et outils de développements utilisés

Les nombreuses recherches effectuées sur la reconnaissance d'entités nommées ont permis le développement de plusieurs outils. Ces derniers présentent une multitude de différences, sur les approches utilisées, sur les langages de programmation associés, sur les types d'entités nommées extraites et sur les langues traitées. Parmi ceux-ci, on peut citer : *Standford NER*[35], *Standford CoreNLP*[36], *Spacy*[37], *LingPipe*[38], *OpenNLP*[39], *Spark NLP*[40] , *cTAKES*[41], *GATE*[42], *etc.* La comparaison est basée sur plusieurs critères tels que l'accessibilité, la facilité d'installation et de la configuration, le langage utilisé, etc. Notre choix est porté sur l'outil *Spacy*, car son installation et sa configuration sont faciles. Il fournit des modèles pré-entrainés et il permet l'entrainement de modèles personnalisés notamment en utilisant une approche à base de règles utilisée dans notre travail.

Nous avons ainsi choisi *Spacy* avec ses méthodes et modèles particulièrement ceux supportant le traitement du français en utilisant le langage de programmation *« Python »* par le biais du Framework *Spyder* de *« Anaconda »*.

- **Spacy** est une bibliothèque python à accès libre qui permet de traiter automatiquement les langues naturelles (tokenisation, POS tagger, NER, etc.). Il dispose de plusieurs modèles pré-entrainés dans plusieurs langues (français, anglais, allemand, espagnol, portugais, etc.). L'outil Spacy est basé sur l'apprentissage statistique en utilisant les CNN[43] (Convolutional Neural Networks, en français réseaux neuronaux convolutifs). Il permet aussi d'entrainer un nouveau modèle selon l'approche souhaitée.
- **Python**[44] est un langage de programmation open source, il ne nécessite pas d'être compilé ; c'est un langage de programmation interprété. Un programme « interpréteur » permet d'exécuter un programme écrit en python sur n'importe quelle autre machine

[35] https://nlp.stanford.edu/software/CRF-NER.html
[36] https://stanfordnlp.github.io/CoreNLP/ner.html
[37] https://spacy.io/
[38] http://www.alias-i.com/lingpipe/
[39] https://opennlp.apache.org/
[40] https://nlp.johnsnowlabs.com/
[41] https://ctakes.apache.org/
[42] https://fabrica.inria.fr/fr/gate/index.html
[43] https://datascientest.com/convolutional-neural-network
[44] https://www.python.org/

(ordinateur), ce qui lui confère la qualité de langage multiplateforme, il est aussi un langage multi paradigme (impérative, orienté-objet, fonctionnelle).

- **Anaconda**[45] est une boite à outils open source pour la science de données en python, R et pour l'apprentissage automatique. Il donne accès à des milliers de packages et de bibliothèques open source.
- **Spyder**[46] est un environnement de travail gratuit, open source et développé en python. Cet outil a été conçu pour les scientifiques, ingénieurs et analystes pour l'édition, l'analyse et l'exploitation des données.

IV.2.4. Configurations

Pour la réalisation de notre modèle d'extraction d'entités nommées, nous avons installé l'utilitaire python et Anaconda. Nous avons aussi installé les bibliothèques python nécessaire notamment celle de spacy ; celle-ci peut s'installer à travers 2 emballages possibles :

- soit avec « *l'emballage pip*[47] » avec la commande suivante :

Illustration 9 : Installation de spacy avec "pip"

- soit avec « *l'emballage conda*[48] » avec la commande suivante :

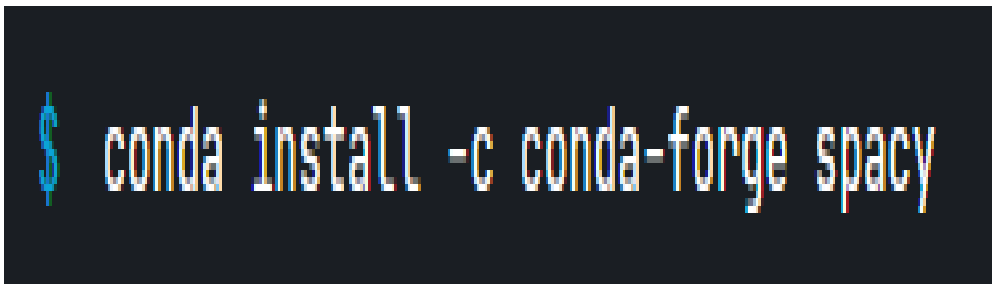

Illustration 10 : Installation de spacy avec "conda"

[45] https://www.anaconda.com/products/individual
[46] https://www.spyder-ide.org/
[47] PIP est un système de gestion de packages utilisé pour installer et gérer des packages logiciels écrit en python.
[48] Conda est l'utilitaire pour installer et gérer les packages installer sur Anaconda.

Apres l'installation de l'outil spacy, il nous a fallu télécharger les modèles des langues (français, anglais, etc.). Spacy propose plusieurs modèles pré-entrainés (à base de CNN « Réseaux de neurones convolutifs ») dans plusieurs langues ; il fournit quatre modèles pour le français et quatre modèles pour l'anglais sous forme de pipeline[49] :

- les modèles pour l'anglais : *« en_core_web_sm », « en_core_web_md », « en_core_web_lg », « en_core_web_trf »* ;
- les modèles pour le français : *« fr_core_news_sm », « fr_core_news_md », « fr_core_news_lg », « fr_dep_news_trf »*.

L'annotation des noms de modèles est spécifiée par quatre attributs :

- la langue du modèle (exemple *« fr »* pour le français et *« en »* pour l'anglais) ;
- la capacité du modèle (les constituants du pipeline) :
  - *Core* (vocabulaire, syntaxe, entité, vecteurs de mot)
  - *Dep* (vocabulaire et syntaxe)
- le type de texte sur lequel il a été entrainé (*« web »* pour l'anglais et *« news »* pour le français) ;
- la taille du texte (« *sm*, small ou petite en français », « *md*, middle ou moyen en français », « *lg*, large ou grande en français » ou bien avec transformateur « *trf* »).

Le téléchargement des modèles de spacy s'effectue avec la commande suivante en spécifiant le modèle choisi (dans cette figure, le modèle anglais « en_core_web_sm » a été choisi) :

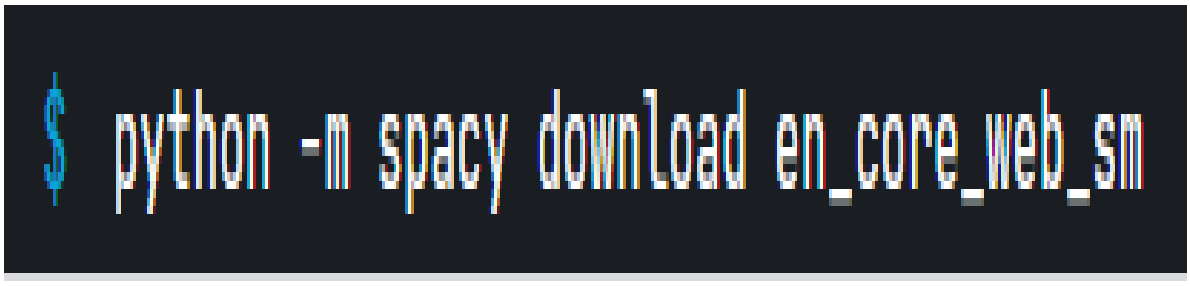

Illustration 11 : Téléchargement du modèle anglais "en_core_web_sm"

Toutes ces étapes sont nécessaires pour une bonne installation des outils à utiliser. Notre méthode utilise une approche à base règles, puisque les entités à extraire sont présentées de

[49] Un pipeline est une chaine de traitement où les instructions sont exécutées en plusieurs étapes

façon plus ou moins uniformes dans les cas cliniques. Ce qui fait qu'elles seront facilement repérables et étiquetables par des règles d'annotations.

Pour cela, nous avons utilisé l'outil « *Matcher* » de *spacy*. Il permet d'extraire des entités nommées en utilisant l'approche à base de règles. Un ensemble de règles[50] sont données en entrée à l'outil « *Matcher* », ce qui va servir de déclencheurs sur lesquels le système va s'appuyer pour repérer le contexte à gauche et à droite pour extraire les mots ou groupes de mots qui sont potentiellement des entités nommées. Ainsi, un ensemble de méthodes (cf IV.1.) ont été utilisées pour l'extraction des différents types d'entités nommées.

IV.2.5. Résultats

Dans cette section, nous allons présenter d'abord des résultats de nos méthodes sur des exemples de cas cliniques. Ensuite, les résultats obtenus sur le corpus de test de la campagne DEFT 2019 seront décrits et comparés avec ceux des participants de cette campagne d'évaluation.

Dans le cas clinique suivant, l'entité de type « genre » est en bleu, l'entité de type « âge » en rouge et leurs catégories respectives en exposant et en vert. Le système extrait ces entités et il les met dans un fichier CSV (le genre sera remplacé par masculin, car ici on parle d'un patient et l'âge sera récupéré directement sans son unité « ans »).

Un patient genre [masculin] de 66 ans âge, traité un an auparavant pour adénocarcinome rectal par amputation abdomino-périnéale avec radiothérapie et chimiothérapie adjuvantes, a été admis pour une urétérohydronéphrose bilatérale sur fibrose retropéritonéale post radique avec insuffisance rénale à 370 mmol/l de créatinine. La montée de sondes double J standard n'a permis qu'une amélioration transitoire de la fonction rénale avec aggravation deux semaines après. La montée de sondes double J trèflées a permis une régression de la dilatation et stabilisation de la fonction rénale à 180 mmol/l pendant 5 mois de suivi.

Illustration 12 : Exemple de cas clinique

[50] Un ensemble de règle fondé sur une base lexicale qui démontre la succession des mots ou tokens pour former ou aider à déterminer une entité.

Le rendu de notre système est récupéré dans le fichier CSV illustré par la figure suivante, le cas concerné est surligné en jaune :

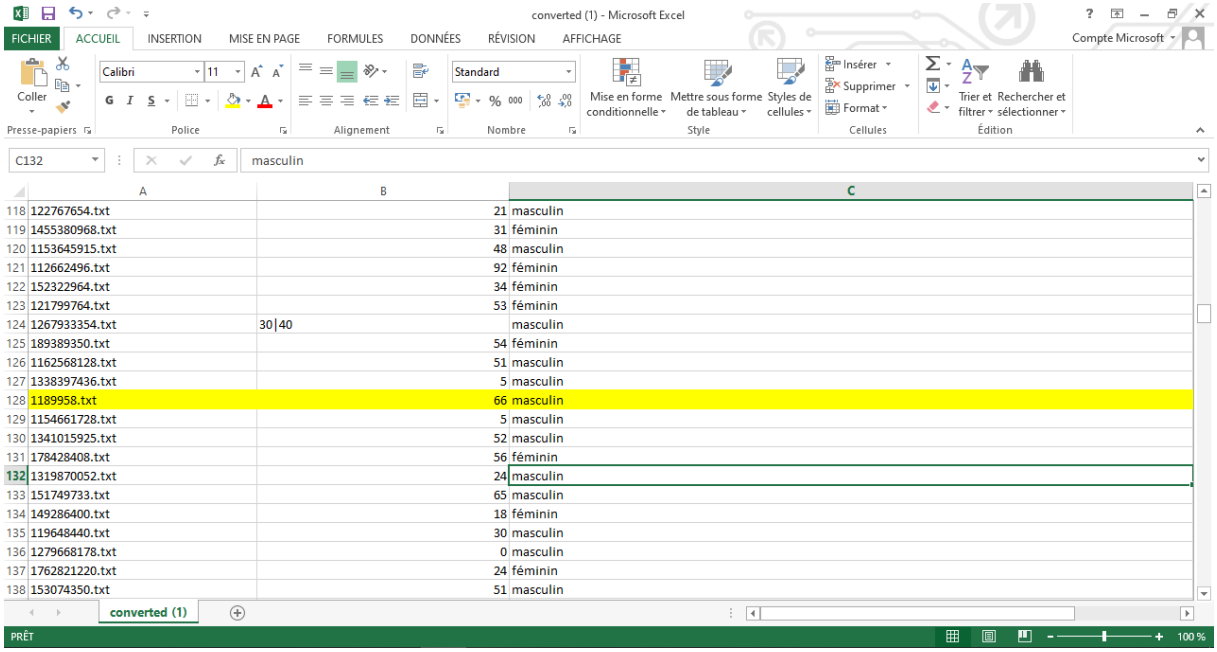

Illustration 13 : Exemple de rendu de notre système

L'évaluation de notre approche à base de règles donne de très bons résultats pour l'extraction des entités de types « *âge* » et « *genre* » sur des cas cliniques avec des f-mesures de plus de 94% pour le type « *âge* » et de 98% pour le type « *genre* ». Notre système est plus efficace pour extraire les entités de type « *genre* » que d'extraire les entités de types « *âge* ». Ceci peut s'expliquer par le fait que plus on augmente le nombre de caractéristiques (features) dans le lexique pour la reconnaissance du genre, plus on couvre l'ensemble des termes identifiant le genre dans les cas cliniques donnés. Alors que pour l'âge, un problème lié à l'ambiguïté sémantique se pose, ce qui réduit la précision. Le tableau 9 présente les résultats obtenus sur le corpus de test de DEFT 2019. Nos règles sont dans l'ensemble assez robustes, mais manquent un certain nombre de motifs.

Types d'EN	Précision	Rappel	f-mesure
Age	0,950	0,935	0,942
Genre	0,977	0,984	0,980

Tableau 14 : Résultats sur le corpus de test de DEFT 2019 pour les EN de types "âge" et "genre"

Le tableau ci-dessous (tableau 10) présente les résultats de notre système nommé **demoNER.** Ils sont comparés à ceux obtenus par les participants à cette campagne, ainsi que les deux baselines[51] (règles et apprentissage (noté ML)), évalués en termes de précision, rappel et f-mesure sur les catégories *« âge »* et *« genre »*. Avec notre système, les âges ont été correctement identifiés dans 94,2% des cas. Le système proposé par LAI donne de meilleurs résultats (94,8%) et le nôtre fournit de meilleurs résultats que les systèmes de EDF Lab (0,624) et de Qwant (0,937). Pour la catégorie du genre, notre approche a obtenu la meilleure f-mesure (98%). Cela peut s'expliquer par la bonne couverture des termes désignant le genre. Le meilleur score pour chaque type d'entités nommées est mis en gras.

Equipes/Systèmes		**demoNER**	**EDF Lab**	**LAI**	**Qwant**	**Baselines**	
						Règles	**ML**
Age	**P**	0,950	0,939	0,980	0,975	0,813	0,961
	R	0,935	0,467	0,919	0,902	0,807	0,912
	F	0,942	0,624	**0,948**	0,937	0,810	0,936
Genre	**P**	0,977	0,967	0,981	0,942	0,934	0,960
	R	0,984	0,472	0,974	0,947	0,928	0,954
	F	**0,980**	0,634	0,978	0,944	0,931	0,957

Tableau 15 : Comparaison des résultats de notre système avec ceux des participants à la campagne DEFT 2019

IV.2.7. Discussion

Notre approche à base de règles a obtenu des résultats satisfaisants sur les données de test de DEFT 2019 avec des F-mesures respectives de plus de 94,2 % et 98% pour la reconnaissance des entités nommées de types « âge » et « genre ». Une analyse de ces résultats a permis d'identifier des problèmes sur la reconnaissance d'entités nommées en dehors des problématiques liées à la délimitation des entités nommées et de leurs catégories respectives. Les limites de notre approche s'expliquent par certains manquements qui sont tous liés à l'approche utilisée, notamment :

- **des règles manquantes :** la plus grande limite de cette approche réside sur le fait qu'il est difficile, voire impossible d'énumérer l'ensemble des termes ou motifs permettant

[51] Les méthodes proposées par les organisateurs

de repérer les entités dans le texte. Ce cas a été rencontré surtout pour la catégorie « âge ». Par exemples : « 26 ans, », « trois mois, »…, ces cas avec la virgule n'ont pas été pris en compte ;

- **la désambiguïsation de l'entité :** en se basant strictement sur les motifs donnés, plusieurs entités peuvent être retrouvées dans un texte. Ainsi, choisir le motif le plus adéquat à extraire nécessite une connaissance sémantique.

  Exemple 1 : *« Un patient de 26 ans décédé à la suite d'un infarctus, a subi à l'âge de 12 ans une transplantation de rein et à 20 ans de foie ».*

  Dans ce cas, on s'attend à ce que le système comprenne à travers les connaissances déjà apprises que l'âge du patient est de 26 ans et non 12 ans ou 20 ans.

  Exemple 2 : *« Le patient qui travaille au sein de la banque ».*

  Dans ce cas, pour déterminer le genre de l'individu concerné, le système doit comprendre que le mot *« sein »* ne correspond pas à l'organe du corps féminin.

Conclusion

Dans ce chapitre, nous avons d'abord présenté la méthodologie utilisée pour extraire les entités nommées de types « âge » et « genre » à partir de cas cliniques. Ensuite, le corpus, les métriques, les outils utilisés, ainsi que les résultats obtenus et une discussion sur ces derniers sont exposés. Enfin, les résultats obtenus ont été dans l'ensemble satisfaisants, mais notre approche présente quelques limites à surmonter pour améliorer ses résultats.

Conclusion générale et perspectives

La reconnaissance d'entités nommées dans le domaine général tout comme dans le domaine biomédical est une tâche majeure et essentielle du traitement automatique du langage naturel. Cette tâche consiste à extraire des unités lexicales spécifiques sur les données textuelles et d'en faire une collection ordonnée. Elle est devenue une étape importante pour plusieurs tâches du TAL, à l'instar de l'extraction d'informations, le résumé automatique et tant d'autres. Ainsi, il existe un ensemble de méthodes, de ressources et d'outils, chacun avec ses avantages et limites. Tout cela laisse entrevoir plusieurs défis à relever pour la mise en place de système de REN robuste et performant.

Dans le cadre de notre mémoire, nous avons effectué une revue de la littérature sur l'extraction d'entités nommées appelée aussi la reconnaissance d'entités nommées (REN) dans le domaine général et dans le domaine biomédical : historique, définitions, approches, métriques, outils, etc. Nous avons ensuite proposé une méthode de reconnaissance des entités nommées de types « âge » et « genre » à partir de données cliniques. L'évaluation de cette méthode sur les corpus (entrainement et test) de la campagne DEFT 2019 a donné de bons résultats.

Ce travail laisse voir plusieurs d'autres pistes de recherches à explorer. En perspectives, nos travaux futurs vont dans le sens d'une part d'améliorer l'approche proposée :

- définir de nouvelles règles manquantes ;
- extraire les autres entités (issue et origine d'admission) ;
- explorer les autres approches notamment l'approche à base d'apprentissage automatique ;
- proposer une plateforme implémentant notre approche.

D'autre part, nous envisageons de concentrer nos recherches sur la création de ressources (corpus annotés) qui sont nécessaires pour créer de tels systèmes et de poursuivre les recherches sur l'extraction des relations entre les entités médicales et le traitement des entités rares et émergentes.

Bibliographie

[1] N. Fourour et E. Morin, « Apport du Web dans la reconnaissance des entités nommées », Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://www.erudit.org/fr/revues/rql/2003-v32-n1-rql1022/012243ar.pdf

[2] L. Besacier, « Des versions imprimées de ces actes peuvent être achetées auprès de »:, p. 41.

[3] Y. Dupont, « La structuration dans les entités nommées », p. 213.

[4] Y. XU, « Reconnaissance d'entités nommées dans les tweets ». Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://er-tim.fr/sites/default/files/XU_Yizhou_2019.pdf

[5] S. Gillot, « Fouille d'opinions », p. 38.

[6] C. Martineau, E. Tolone, et S. Voyatzi, « Les Entités Nommées : usage et degrés de précision et de désambiguïsation », in *26ème Colloque international sur le Lexique et la Grammaire (LGC'07)*, Bonifacio, France, oct. 2007, p. pages 105-112. Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://hal.archives-ouvertes.fr/hal-00461891

[7] M. Ehrmann, « Les Entitées Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation », Theses, Paris Diderot University, 2008. Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://hal.archives-ouvertes.fr/tel-01639190

[8] T. Poibeau, « Le traitement automatique des langues pour les sciences sociales : quelques éléments de réflexion à partir d'expériences récentes », *Réseaux*, vol. 2014/6, n° 188, p. 25-51, 2014, doi: 10.3917/res.188.0025.

[9] R. Grishman et B. Sundheim, « Message Understanding Conference- 6: A Brief History », 1996. doi: 10.3115/992628.992709.

[10] S. T. Aguilar, « Un modèle de reconnaissance automatique des entités nommées et des structures textuelles pour les corpus diplomatiques médiolatins. », p. 305.

[11] X. Tannier, « Analyse de Textes et Extraction d'Information », p. 77.

[12] L. Kheira, « Les annotations sémantiques dans les documents Web : application aux textes psychologiques en langue arabe ». Consulté le: 18 mai 2022. [En ligne]. Disponible sur: http://www.univ-usto.dz/theses_en_ligne/doc_num.php?explnum_id=2784

[13] H. SOUIAH, « Reconnaissance d'entités nommées par une approche bioinspirée ».

[14] D. NOUVEL, « RECONNAISSANCE DES ENTITÉS NOMMÉES PAR EXPLORATION DE RÈGLES D'ANNOTATION Interpréter les marqueurs d'annotation comme instructions de structuration locale ». Consulté le: 18 mai 2022. [En ligne]. Disponible sur: http://www.applis.univ-tours.fr/theses/2012/damien.nouvel_4264.pdf

[15] N. FRIBURGER, « Reconnaissance automatique des noms propres application a la classification de textes journalistiques ».

[16] C. Grouin, O. Galibert, S. Rosset, L. Quintard, et P. Zweigenbaum, « Mesures d'évaluation pour entités nommées structurées », p. 14.

[17] M. Hatmi, « Adaptation d'un système de reconnaissance d'entités nommées pour le français à l'anglais à moindre coût », in *RECITAL*, Grenoble, France, juin 2012, vol. 3, p. 151-161. Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://hal.archives-ouvertes.fr/hal-00727779

[18] E. F. Tjong Kim Sang et F. De Meulder, « Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition », in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 2003, p. 142-147. Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://aclanthology.org/W03-0419

[19] A. Fotsoh, « Recherche d'entités nommées complexes sur le Web - propositions pour l'extraction et pour le calcul de similarité », Theses, Université de Pau et des Pays de l'Adour, 2018. Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://hal.archives-ouvertes.fr/tel-02414263

[20] F. Ben Mesmia, « Reconnaissance des entités nommées à partir de Wikipédia arabe », Theses, Université de Tunis El Manar, 2019. Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://hal.archives-ouvertes.fr/tel-03325717

[21] D. Nouvel, « Reconnaissance des entités nommées par exploration de règles d'annotation - Interpréter les marqueurs d'annotation comme instructions de structuration locale », p. 180.

[22] C. Gizard, « Conception d'une m´ethode hybride d'extraction d'informations g´eographiques `a partir de donn´ees textuelles ». Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://imu.universite-lyon.fr/wp-content/uploads/2018/05/GizardChristopher.pdf

[23] F. DEFFAF, « EXTRACTION DES ENTITÉS NOMMÉES PAR PROJECTION CROSSLINGUISTIQUE ET CONSTRUCTION DE LEXIQUES BILINGUES D'ENTITÉS NOMMÉES POUR LA TRADUCTION AUTOMATIQUE STATISTIQUE ». Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://archipel.uqam.ca/7540/1/M13777.pdf

[24] R. Weegar, A. Pérez, A. Casillas, et M. Oronoz, « Recent advances in Swedish and Spanish medical entity recognition in clinical texts using deep neural approaches », *BMC*

Med. Inform. Decis. Mak., vol. 19, nº 7, p. 274, déc. 2019, doi: 10.1186/s12911-019-0981-y.

[25] E. Kogkitsidou, « Extraction automatique d'entités toponymiques et noms communs liés à la ville », p. 42.

[26] M. Hatmi, « Reconnaissance des entités nommées dans des documents multimodaux », Theses, UNIVERSITÉ DE NANTES, 2014. Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://hal.archives-ouvertes.fr/tel-01154811

[27] B. Sagot, M. Richard, et R. Stern, « Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées », p. 9.

[28] R. Rakotomalala, « Opinion mining and sentiment analysis ». Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://eric.univ-lyon2.fr/~ricco/cours/slides/WM.A%20-%20Opinion%20mining%20and%20sentiment%20analysis.pdf

[29] F. Béchet, B. Sagot, et R. Stern, « Coopération de méthodes statistiques et symboliques pour l'adaptation non-supervisée d'un système d'étiquetage en entités nommées », p. 7.

[30] I. Savard, « Exploration d'articles scientifiques sur les maladies rares pour l'extraction d'informations », p. 93.

[31] D. Maurel, N. Friburger, J.-Y. Antoine, I. Eshkol, et D. Nouvel, « Cascades de transducteurs autour de la reconnaissance des entités nommées », *Rev. TAL*, vol. 52, nº 1, p. 69-96, 2011.

[32] D. Nouvel, J.-Y. Antoine, N. Friburger, et A. Soulet, « Fouille de règles d'annotation pour la reconnaissance d'entités nommées », *Rev. TAL*, vol. 54, nº 2, p. 13-41, 2013.

[33] N. Okinina, D. Nouvel, N. Friburger, et J.-Y. Antoine, « Apprentissage supervisé sur ressources encyclopédiques pour l'enrichissement d'un lexique de noms propres destiné à la reconnaissance des entités nommées », in *TALN'2013, 20e conférence sur le Traitement Automatique des Langues Naturelles*, Les Sable d'Olonne, France, juin 2011, p. 667-674. Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://hal.archives-ouvertes.fr/hal-01016545

[34] C. JIANG, « Extraction d'Entités d'Aliments/Médicaments à Partir de Textes Biomédicaux en Français ». Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://er-tim.fr/sites/default/files/JIANG_Chunyang_2019.pdf

[35] W. Hafiane, « Apprentissage par transfert pour l'extraction de relations pharmacogénomiques à partir de textes », p. 78.

[36] S. Wu *et al.*, « Deep learning in clinical natural language processing: a methodical review », *J. Am. Med. Inform. Assoc. JAMIA*, vol. 27, nº 3, p. 457-470, déc. 2019, doi: 10.1093/jamia/ocz200.

[37] J. P. C. Chiu et E. Nichols, « Named Entity Recognition with Bidirectional LSTM-CNNs ». arXiv, 19 juillet 2016. Consulté le: 18 mai 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1511.08308

[38] Z. Huang, W. Xu, et K. Yu, « Bidirectional LSTM-CRF Models for Sequence Tagging ». arXiv, 9 août 2015. Consulté le: 18 mai 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1508.01991

[39] J. Devlin, M.-W. Chang, K. Lee, et K. Toutanova, « BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding ». arXiv, 24 mai 2019. Consulté le: 18 mai 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1810.04805

[40] M. D. T. Nzali, A. Névéol, et X. Tannier, « Analyse d'expressions temporelles dans les dossiers électroniques patients », in *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, Caen, France, juin 2015, p. 49-58. Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://aclanthology.org/2015.jeptalnrecital-long.5

[41] V. Claveau, « Extraction d'informations (domaine biomédical) », p. 10.

[42] I. Zribi, S. Mezghani Hammami, et L. Hadrich Belguith, « L'apport d'une approche hybride pour la reconnaissance des entités nommées en langue arabe », in *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, Montréal, Canada, juill. 2010, p. 183-188. Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://aclanthology.org/2010.jeptalnrecital-court.31

[43] J. Makhoul, F. Kubala, R. Schwartz, et R. Weischedel, « PERFORMANCE MEASURES FOR INFORMATION EXTRACTION », p. 4.

[44] O. Galibert *et al.*, « Named and specific entity detection in varied data: the Quaero named entity baseline evaluation », p. 7.

[45] M. Jannet, M. Adda-Decker, O. Galibert, J. Kahn, et S. Rosset, « ETER: a New Metric for the Evaluation of Hierarchical Named Entity Recognition », mai 2014.

[46] N. Friburger, « Linguistique et reconnaissance automatique des noms propres », *Meta J. Trad. Meta Transl. J.*, vol. 51, nº 4, p. 637-650, 2006, doi: 10.7202/014331ar.

[47] A. GHOULAM, « L'extraction d'information pour la recherche dans un système médical a large échelle ». Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://theses.univ-oran1.dz/document/TH4924.pdf

[48] S. Paumier, « Unitex-GramLab-3.1-usermanual-fr.pdf ». Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://unitexgramlab.org/releases/3.1/man/Unitex-GramLab-3.1-usermanual-fr.pdf

[49] N. Fourour, « Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français », in *Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, Nancy, France, juin 2002, p. 267-276. Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://aclanthology.org/2002.jeptalnrecital-long.24

[50] A.-L. Minard, A. Roques, N. Hiot, M. Halfeld Ferrari Alves, et A. Savary, « DOING@DEFT : cascade de CRF pour l'annotation d'entités cliniques imbriquées », in *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, Nancy, France, 2020, p. 66-78. Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://hal.archives-ouvertes.fr/hal-02784743

[51] POTIER, « Fouille de résumés d'articles biomédicaux Annotation d'entités et extraction de relations ». Consulté le: 18 mai 2022. [En ligne]. Disponible sur: http://www.tal.univ-paris3.fr/plurital/memoires/juliette-potier-16-17.pdf

[52] D. Demner-Fushman, W. W. Chapman, et C. J. McDonald, « What can Natural Language Processing do for Clinical Decision Support? », *J. Biomed. Inform.*, vol. 42, nº 5, p. 760-772, oct. 2009, doi: 10.1016/j.jbi.2009.08.007.

[53] S. Atdağ et V. Labatut, « A Comparison of Named Entity Recognition Tools Applied to Biographical Texts », in *2nd International Conference on Systems and Computer Science*, Villeneuve d'Ascq, France, août 2013, p. 6p. Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://hal.archives-ouvertes.fr/hal-00849797

[54] C. Raymond et J. Fayolle, « Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement », p. 11.

[55] Breckbaldwin, « Coding Chunkers as Taggers: IO, BIO, BMEWO, and BMEWO+ », *LingPipe Blog*, 14 octobre 2009. https://lingpipe-blog.com/2009/10/14/coding-chunkers-as-taggers-io-bio-bmewo-and-bmewo/ (consulté le 18 mai 2022).

[56] P. Prakash, « Extend Named Entity Recogniser (NER) to label new entities with spaCy », *Medium*, 18 septembre 2020. https://towardsdatascience.com/extend-named-entity-recogniser-ner-to-label-new-entities-with-spacy-339ee5979044 (consulté le 18 mai 2022).

[57] C. Grouin, « Anonymisation de documents cliniques: performances et limites des méthodes symboliques et par apprentissage statistique », p. 249.

[58] A. Névéol, C. Grouin, J. Leixa, S. Rosset, et P. Zweigenbaum, « The Quaero French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization », p. 7.

[59] I. Pérez-Díez, R. Pérez-Moraga, A. López-Cerdán, J.-M. Salinas-Serrano, et M. de la Iglesia-Vayá, « De-identifying Spanish medical texts - named entity recognition applied to radiology reports », *J. Biomed. Semant.*, vol. 12, nº 1, p. 6, mars 2021, doi: 10.1186/s13326-021-00236-2.

[60] A. Bouffier, C. Duclos, et T. Poibeau, « Analyse et structuration automatique des guides de bonnes pratiques cliniques : essai d'évaluation. », in *19es Journées Francophones d'Ingénierie des Connaissances (IC 2008)*, Nancy, France, juin 2008, p. 135-146. Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://hal.archives-ouvertes.fr/hal-00416694

[61] A. Ben Abacha et P. Zweigenbaum, « Une étude comparative empirique sur la reconnaissance des entités médicales », *TAL J. Trait. Autom. Lang.*, vol. 53, p. 39-68, janv. 2012.

[62] V. Kumar, A. Stubbs, S. Shaw, et Ö. Uzuner, « Creation of a new longitudinal corpus of clinical narratives », *J. Biomed. Inform.*, vol. 58, p. S6-S10, déc. 2015, doi: 10.1016/j.jbi.2015.09.018.

[63] L. Deléger et A. Névéol, « Automatic identification of document sections for designing a French clinical corpus (Identification automatique de zones dans des documents pour la constitution d'un corpus médical en français) [in French] », in *Proceedings of TALN 2014 (Volume 2: Short Papers)*, Marseille, France, juill. 2014, p. 568-573. Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://aclanthology.org/F14-2030

[64] C. Grouin et A. Névéol, « De-identification of clinical notes in French: towards a protocol for reference corpus development », *J. Biomed. Inform.*, vol. 50, p. 151-161, août 2014, doi: 10.1016/j.jbi.2013.12.014.

[65] A. Névéol, « Traitement Automatique de la Langue Biomédicale », Habilitation à diriger des recherches, Université Paris Sud, 2018. Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://hal.archives-ouvertes.fr/tel-02167096

[66] G. K. Savova *et al.*, « Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications », *J. Am. Med. Inform. Assoc.*, vol. 17, nº 5, p. 507-513, sept. 2010, doi: 10.1136/jamia.2009.001560.

[67] C. Dalloux, N. Grabar, et V. Claveau, « Détection de la négation : corpus français et apprentissage supervisé », *Rev. Sci. Technol. Inf. - Sér. TSI Tech. Sci. Inform.*, p. 1-21, déc. 2019.

[68] C. Grouin, « Guide d'annotation des effets secondaires rapportés par les patients sur les réseaux sociaux », p. 16.

[69] L. Campillos, L. Deléger, C. Grouin, T. Hamon, A.-L. Ligozat, et A. Névéol, « A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT) », *Lang. Resour. Eval.*, vol. 52, nº 2, p. 571-601, juin 2018, doi: 10.1007/s10579-017-9382-y.

[70] C. Grouin, N. Grabar, V. Claveau, et T. Hamon, « Clinical Case Reports for NLP », in *BioNLP 2019 - 18th ACL Workshop on Biomedical Natural Language Processing*, Florence, Italy, août 2019, p. 273-282. doi: 10.18653/v1/W19-5029.

[71] A. Ben Abacha et P. Zweigenbaum, « Automatic extraction of semantic relations between medical entities: a rule based approach », *J. Biomed. Semant.*, vol. 2, nº 5, p. S4, oct. 2011, doi: 10.1186/2041-1480-2-S5-S4.

[72] A. Ben Abacha, P. Zweigenbaum, et A. Max, *Extraction d'information automatique en domaine médical par projection inter-langue : vers un passage à l'échelle*. 2012.

[73] N. Grabar, C. Grouin, T. Hamon, et V. Claveau, « Recherche et extraction d'information dans des cas cliniques. Présentation de la campagne d'évaluation DEFT 2019 », in *DEFT 2019 - Défi fouille de texte*, Toulouse, France, juill. 2019, p. 1-10. Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://hal.archives-ouvertes.fr/hal-02280852

[74] R. Cardon, N. Grabar, C. Grouin, et T. Hamon, « Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques ». Consulté le: 18 mai 2022. [En ligne]. Disponible sur: https://hal.archives-ouvertes.fr/hal-02784737