Institut National des Langues et Civilisations Orientales

Département Textes, Informatique, Multilinguisme

Extraction d'Entités d'Aliments/Médicaments à Partir de Textes Biomédicaux en Français

MASTER

TRAITEMENT AUTOMATIQUE DES LANGUES

Parcours :

Ingénierie Multilingue

par

Chunyang JIANG

Directeur de mémoire :

Mathieu Valette

Encadrant :

Antonio Balvet

Année universitaire 2018/2019

TABLE DES MATIÈRES

Liste des figures

Liste des tableaux

INTRODUCTION

Contexte et Motivations

On pense généralement aux interactions entre médicaments (interactions médicamenteuses). Cependant, des interactions peuvent également exister entre les médicaments et les aliments (interactions aliments-médicaments), les médicaments et les nutriments (interactions nutriments-médicaments), ainsi qu'entre les médicaments et les herbes (interactions plantes-médicaments)[Sørensen, 2002]. L'effet d'un médicament sur une personne pourrait être différent de celui attendu, car ce médicament interagit avec un autre médicament pris par la personne (interactions médicamenteuses), avec des aliments, des boissons, des compléments alimentaires que la personne consomme (interactions aliments-médicaments) ou une autre maladie que la personne a (interactions médicament-maladies) [Sørensen, 2002, Bushra et al., 2011]. D'après [Frankel, 2003, Bista et al., 2006], une interaction médicamenteuse est une situation dans laquelle une substance affecte l'activité d'un médicament, c'est-à-dire que ses effets sont augmentés ou diminués, ou qu'ils produisent un nouvel effet qui ne produit pas par lui-même. Les interactions médicamenteuses peuvent modifier la pharmacocinétique (l'ADME) et/ou la pharmacodynamique d'un médicament [Palleria et al., 2013]. Ces interactions, qui peuvent résulter d'une mauvaise utilisation accidentelle ou d'un manque de connaissances sur les ingrédients actifs impliqués dans les substances concernées, représentent une source d'erreur de médication importante et largement méconnue [Bista et al., 2006].

L'étude des interactions médicamenteuses devrait améliorer la sécurité des médicaments et permettre un traitement médicamenteux personnalisé [Bista et al., 2006]. Il est important de connaître les interactions médicamenteuses possibles et de les communiquer clairement aux patients, aux médicaments et aux pharmaciens. Pour cette raison, depuis plusieurs années, l'extraction d'informations sur les entités médicamenteuses et les interactions médicamenteuses fait l'objet de nombreuses recherches, compétitions et campagnes d'évaluation, par exemple dans [Segura-Bedmar et al., 2013, Björne et al., 2013, Ben Abacha et al., 2015a, Patel and Beckett, 2016, Wang et al., 2017, Fung et al., 2017, Lim et al., 2018, Sun et al., 2018, Xu et al., 2018, Zhang et al., 2019], afin de fournir des informations actualisées et à jour sur les médicaments ou sur les interactions connues entre médicaments ou entre leurs principes actifs [Hamon et al., 2017].

Cependant, la reconnaissance des entités d'aliments ne rarement fait l'objet des recherches dans le domaine de la fouille de texte biomédical [Huang and lu, 2015, Hamon et al., 2017]. Extraire les mentions de médicaments et d'aliments dans les documents biomédicaux est une étape importante voire élémentaire pour permettre une gamme d'applications de fouille de textes en aval, telles que la détection des in-

teractions médicamenteuses [Segura-Bedmar et al., 2013], des interactions aliments-médicaments [Hamon et al., 2017, Bordea et al., 2018, Randriatsitohaina, 2018], ainsi que les effets indésirables médicamenteuses (ADRs)[Oronoz et al., 2015] qui en résultent.

Problématique et Objectif

Contrairement à l'accès facile à l'information sur les médicaments et les interactions médicamenteuses, ainsi qu'aux efforts de recherche intensifs dans ce domaine, l'information sur les interactions aliments-médicaments n'est pas toujours facilement disponible, sans parler du manque de recherche pertinent. En fait, la plupart des informations sur les interactions aliments-médicaments et les effets indésirables associés sont actuellement fragmentées et dispersées dans des sources hétérogènes, telles que des articles scientifiques et certaines bases de connaissances, telles que DrugBank[Wishart et al., 2006], ou éventuellement dans des forums de discussion fournissant le point de vue d'un patient sur les événements indésirables[Hamon et al., 2017]. En outre, les informations sont principalement disponibles en anglais, tandis que les connaissances pertinentes se trouvent également dans des données textuelles écrites en français. Pourtant, cette information reste pauvre[Hamon et al., 2017, Névéol et al., 2015, Névéol et al., 2014].

À cet égard, il est important de rassembler et d'utiliser ces ressources, structurées ou non, afin d'aider à formaliser et à visualiser la description des aliments, des médicaments et des interactions aliments-médicaments. Il est donc envisageable de surmonter ce problème en profitant des méthodes et des techniques appuyant sur la fouille de textes et sur le traitement automatique des langues [Cohen and Hunter, 2008, Zweigenbaum et al., 2007].

L'objectif est naturellement la constitution d'un corpus français sur les interactions aliments-médicaments pour pouvoir ensuite proposer des approches, symboliques ou statistiques, permettant l'extraction des entités aliments/médicaments, étape fondamentale pour l'extraction des interactions, l'alimentation d'une base de connaissances, ainsi que pour les autres tâches en aval.

Projet MIAM

Ce mémoire s'inscrit dans le cadre du stage de fin d'études dans le Projet de recherches ANR MIAM [1] (Maladies, Interactions Alimentation-Médicaments), cotravaillé par

- LIMSI (UPR3251) - Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur - CNRS, Université Paris-Sud
- STL (UMR 8163 CNRS) - Savoirs, textes, langage - Université de Lille - sciences humaines et sociales
- Université de Bordeaux
- CNHIM (Association) Centre National Hospitalier d'Information sur le Médicament
- ANTIDOT (Recherche Développement)

1. https://miam.limsi.fr/index.php

Avec le volume important de données non structurées présentes dans les bases bibliographiques, mais aussi le développement de bases de connaissances ouvertes, accéder aux connaissances qu'elles contiennent nécessitent de produire une vision globale à partir de multiples sources hétérogènes. Pour cela, le projet MIAM vise à proposer des méthodes s'appuyant sur le Traitement Automatique des Langues, la fouille de textes, ainsi que la représentation et la modélisation de connaissances, afin d'agréger ces données et connaissances issues de bases de connaissances, de Linked Open Data ou de publications scientifiques. L'évaluation des résultats est réalisée avec cas d'usage réel : les interactions entre des médicaments et des aliments pouvant conduire à un effet indésirable. Ces informations sont généralement dispersées dans plusieurs ressources et les agréger aidera à formaliser et à visualiser la description de ces interactions pour éviter ce genre d'effets.

Organisation du mémoire

Ce travail est organisé en 4 chapitres.

— Le Chapitre 1 donne un aperçu des concepts théoriques qui constituent la base de ce mémoire. Après une introduction à la fouille de textes, les méthodes de reconnaissance d'entités nommées sont décrites ; il présente des recherches pertinentes, d'approches similaires et de ressources utiles. Les informations sur les travaux connexes fournissent des indications pour la sélection des méthodes les mieux adaptées à notre tâche ;

— Le Chapitre 2, en s'appuyant sur les idées inspirées du Chapitre 1, porte sur la méthodologie et le matériel pour ce travail ;

— La configuration expérimentale est présentée au Chapitre 3. Un corpus de test est créé pour évaluer les méthodes mises en œuvre ;

— Dans le Chapitre 4, des mesures de performance sont calculées et les résultats sont discutés.

Première partie

Contexte général

CHAPITRE 1

CONTEXTE THÉORIQUE

Sommaire

1.1 Introduction

Dans ce chapitre, les fondements théoriques de ce domaine de recherche sont expliqués. Les concepts de base ainsi que des approches plus spécifiques sont décrits à ce stade. Les théories et les concepts présentés dans ce chapitre sont ensuite mis en pratique lors de la conception du système de reconnaissance des entités aliments/médicaments.

1.2 Fouille de textes

[Zweigenbaum et al., 2007] conclut qu'il existe des définitions plus larges et plus strictes de la fouille de textes. Selon la définition la plus stricte du terme, un système de fouille de textes doit renvoyer des connaissances qui ne sont pas explicitement énoncées dans les textes [Hearst, 2003]. Selon cette définition, certains systèmes de résumé et de questions-réponses seraient qualifiés de fouille de textes. Sur une définition plus large, un système de fouille de textes extrait des informations à partir des textes ou exécute des fonctions indispensables à cette opération [Jackson, 2003]. Les systèmes ayant cet objectif sont couramment appelés applications d'extraction d'information, à savoir : reconnaissance d'entité nommée, extraction de relation, extraction d'événement, etc. De tels systèmes effectuent généralement la reconnaissance d'entité nommée comme étape de traitement initiale.

La fouille de textes du domaine biomédical (*biomedical text mining*), également connu sous le nom de *BioNLP* ou traitement de langage biomédical, est

l'application de techniques de traitement automatique des langues naturelles aux données biomédicales [Cohen, 2010]. La plupart des recherches reposent, à des degrés divers, sur des méthodes et des outils de traitement du langage naturel [Zweigenbaum et al., 2007, Cohen, 2010]. En pratique, *BioNLP* et la fouille de textes du domaine biomédical se situent sur un continuum. Dans ce travail, la fouille de textes du domaine biomédical est la désignation que nous utilisons ci-après.

Certaines recherches [Ananiadou et al., 2005, Zweigenbaum et al., 2007, Chapman and Cohen, 2009, Karsten and Suominen, 2009, Cohen, 2010] indiquent que les outils de fouille de textes du domaine général ne sont pas bien adaptés au domaine biomédical, en raison de sa nature hautement spécialisée. Malgré ce fait, les tâches partagent un mode de traitement commun, à savoir le traitement d'un grand volume de documents d'entrée sous forme de *pipeline* [Cohen, 2014, Névéol, 2018], comme le montre le digramme 1.1.

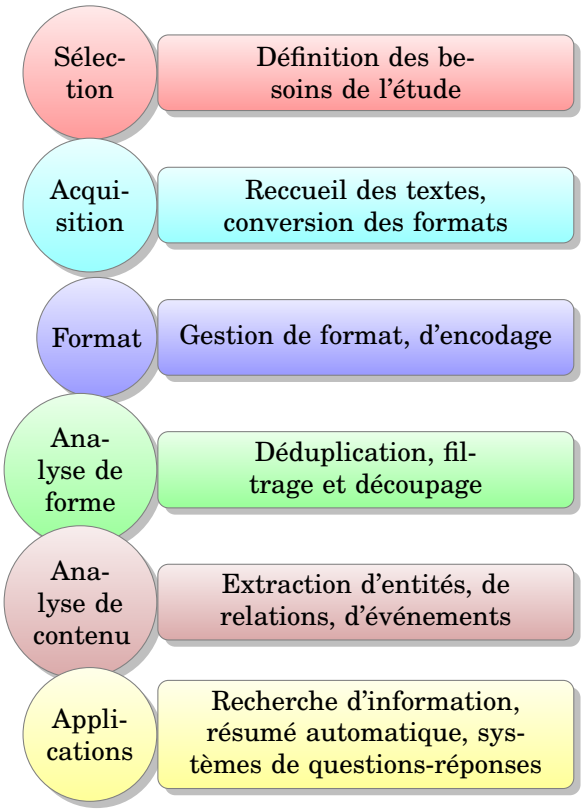


FIGURE 1.1 – Chaîne de traitements de la fouille de textes

Des aspects de la langue du domaine biomédical pourraient en réalité rendre le traitement plus difficile que les textes du domaine général, malgré le fait que ses textes soient sémantiquement plus restreints ; pourtant, il existe des ressources, lexicales et autres, qui pourraient rendre les textes biomédicaux plus facile à traiter que les textes du domaine général [Cohen, 2010]. Dans le domaine général, il n'y a pas d'intérêt dominant sur un seul genre. Les genres fréquemment extraits incluent :

— nouvelles ;
— encyclopédie, en particulier Wikipédia ;
— microblogs, en particulier les tweets ;

— journaux de requêtes ;
— fictions ;
— transcriptions de conversation.

Alors que dans le domaine biomédical, la littérature scientifique est le genre dominant. Les autres genres souvent exploités comprennent :
— portails de santé encyclopédiques sur le Web ;
— forums de discussion des patients sur le Web ;
— nouvelles ;
— microblogs, en particulier les tweets ;
— emballage de médicaments ;
— textes cliniques tels que les dossiers des patients et les transcriptions cliniques ;
— documents d'essais cliniques ;
— requêtes des patients ;
— brevets.

1.3 Reconnaissance d'entités nommées

Le terme «entité nommée» (désormais EN) a été utilisé pour la première fois en 1995 lors de la sixième Conférence de Compréhension du Message (MUC-6) [Grishman and Sundheim, 1996], en tant que tâche d'identification des noms d'organisations, de personnes et de lieux géographiques dans un texte, ainsi que des expressions de monnaie, de temps et de pourcentage. Dans cette sous-section, nous nous concentrons principalement sur les définitions, les étapes principales, les approches et les méthodes d'évaluation des tâches de la reconnaissance d'entités nommées(désormais REN).

1.3.1 Définition et principes

Dans ce travail, nous n'entrerons pas dans les détails concernant la définition du sens réel des entités nommées, car cela pose un défi philosophique, qui est par exemple discuté dans [Nouvel et al., 2016]. Par conséquent, dans le contexte de nos travaux, une définition plutôt pragmatique exprimant le rôle des entités est utilisée.

Définition 1.1. Entité nommée : Une entité ou une entité nommée est une mention (formes de surface telles que des phrases nominales, des expressions simples ou multiples, etc.) qui fait référence à un objet identique du monde réel ou à un concept abstrait [Ernst, 2017].

Exemple : Les mentions de *Institut national des langues et civilisations orientales*, *INaLCO*, *Langue O'*, font référence à la même entité de type `organisation`.

Définition 1.2. Entité biomédicale : Une entité biomédicale ou bien entité nommée biomédicale désigne une instance d'un concept sémantiquement bien définie dans le domaine biomédical [Leser and Hakenberg, 2005, Simpson and Demner-Fushman, 2012].

Les entités d'intérêt communes comprennent les noms de gènes et de protéines, les problèmes et traitements médicaux, les noms de médicaments et leurs dosages

[Simpson and Demner-Fushman, 2012], etc.

Exemple : De même, *cholécalciférol* et *Vitamine D3* sont deux formes de surface désignées de la même entité de type `vitamine` (un groupe thérapeutique de la classe ATC [1] A11).

Pour catégoriser des entités homogènes en classes, des types sont introduits :

Définition 1.3. Type d'entité : Un type d'entité désigne un ensemble d'entités qui partagent une sémantique et des caractéristiques communes. Nous définissons l'ensemble de tous les types $T = \{t_n | n \in \mathbb{N}\}$.

Exemple : Les entités *ER-TIM*, *INaLCO*, *CNRS* sont toutes membres des types `organisation` et `institution académique`, alors que *Vitamine D3*, *Vitamine B2* ou *riboflavine* sont membres du type `vitamine`.

Définition 1.4. Reconnaissance d'entités nommées : La reconnaissance d'entités nommées (REN) est le processus de localisation et de catégorisation d'entités nommées dans des textes en types d'entités prédéfinis.

Sur cette base, la REN pourrait naturellement être défini comme deux sous-tâches interdépendantes, comme indiqué dans [Nouvel et al., 2016] :

— **Détection :** déterminer les segments concernés, c'est-à-dire le début et la fin des entités nommées (index de séquence) ;
— **Catégorisation :** identifier la catégorie (*i.e.* le type) d'entités nommées contenues dans chaque segment détecté à l'aide d'une liste prédéfinie (`personne`, `lieu`, `date`, etc.).

Formellement, étant donné une séquence de mots $s = \langle w_1, w_2, \ldots, w_n \rangle$, un système REN doit générer une liste de triplets $\langle i_s, i_e, t_i \rangle$, chacun d'eux étant une entité nommée mentionnée dans s. Ici, $i_s \in [1, n]$ et $i_e \in [1, n]$ sont les indices de début et de fin d'une entité nommée, $i_s \leqslant i_e$; t est le type d'entité d'un groupe de types prédéfini, $t \in T$ tel que défini dans la définition 1.3, d'après [Li et al., 2018].

Exemple : Étant donné la phrase « Le siège de l'INLCO se trouve au cœur du nouveau quartier latin ». Un système REN serait en mesure de reconnaître que les termes *INaLCO* et *Nouveau quartier latin* sont des mentions d'entités et appartiennent respectivement à une `organisation` et à un `lieu`.

Exemple : Étant donné la phrase « Le brocoli est indispensable entre la coagulation sanguine et la vitamine K ». *brocoli*, *vitamine K* et *coagulation sanguine* seraient identifiés et classés comme `aliment`, `composant alimentaire` et `effet des médicaments`.

Il faut noter que l'identification et la catégorisation peuvent être effectuées de manière consécutive ou simultanée [Ben Abacha and Zweigenbaum, 2011].

1. Classification anatomique, thérapeutique et chimique, développée par l'OMS (Organisation mondiale de la Santé) pour classer les médicaments et autres produits médicaux. Voir. https://www.whocc.no/atc/structure_and_principles/

1.3.2 Typologie

Les types les plus étudiés pour les tâches de reconnaissance d'entités biomédicales sont les gènes, les protéines, les produits chimiques, les médicaments, les maladies, les symptômes, etc. Les annotations en entités biomédicales souvent s'appuient sur les concepts de l'UMLS [2], tel est le cas du corpus QUAERO Médical du français [3] [Névéol et al., 2014]. Certes, il existe bien d'autres typologies de concepts biomédicales en fonction de sous-domaine étudié et de tâches envisagées.

Dans le cadre du projet MIAM, les experts ont développé un guide d'annotation [4] pour annoter un corpus en anglais, qui est composé des titres et résumés d'articles scientifiques extraits du Medline. Tenant en compte des matériels disponibles en français, dans ce mémoire, nous nous concentrons sur 4 types d'entités, à savoir `food`, `component`, `drug`, `drugClass` :

- **`food`** : aliments de base ou transformés, y compris les suppléments alimentaires ;
  - *aliment*, *boisson*, *lait*, *viande*, *fruit*, *eau*, *brocoli*, ...
  - *alcool*, *jus d'orange*, *chocolat*, ...
  - *ginsin*, *goji*, *lavande*, ...
- **`component`** : composants alimentaires qui ne sont pas pris comme compléments ou suppléments alimentaires ;
  - *alcool*, *caféine*, *vitamine K*, *vitamines A*, ...
  - Certains composants sont sources d'ambiguïté : *vitamine* est un apport comme médicament mais sera plus souvent « consommé » comme composant alimentaire. La désambiguïsation de type se fait en fonction du contexte.
- **`drug`** : médicaments , leurs métabolites et les isomères + prodrug, enantomiers + active/inactive metabolite ;
  - nom commercial, générique, DCI [5] ;
  - *simvastatine*, *warfarine*, ...
- **`drugClass`** : classes pharmacologiques des médicaments, type de thérapie suivie.
  - *anticestodes*, *stéroïdes anabolisants*, ...
  - *thérapeutique hépatique et biliaire*, *thérapeutique antivariqueuse*, ...

1.3.3 Méthodes

En ce qui concerne les approches REN, différentes recherches peuvent utiliser différentes taxonomies, par exemple, comme indiqué dans les recherches [Poibeau, 2003, Ben Abacha and Zweigenbaum, 2011, Simpson and Demner-Fushman, 2012, Liu et al., 2015, Nouvel et al., 2016]. Dans cette sous-section, nous allons tracer les grandes lignes de ses méthodes en 4 catégo-

2. https://www.nlm.nih.gov/research/umls/index.html
3. https://quaerofrenchmed.limsi.fr/
4. *N.B.* L'auteur du mémoire n'est pas propriétaire du guide d'annotation ; il sera publié ultérieurement par le groupe de recherche du Projet MIAM.
4. Base de données bibliographiques regroupant la littérature relative aux sciences biologiques et biomédicales. Voir. https://www.ncbi.nlm.nih.gov/pubmed/
5. Dénominations communes internationales. Voir. https://www.who.int/medicines/services/inn/innguidance/fr/

ries : à base de dictionnaire, à base de règles, à base d'apprentissage automatique et hybrides [Simpson and Demner-Fushman, 2012, Liu et al., 2015].

1.3.3.1 À base de dictionnaire

Définition 1.5. Dictionnaire/Lexique : [Nouvel et al., 2016] indique,

> « [..] dans le contexte spécifique de la reconnaissance d'entités nommées, la notion de lexique fait référence à la plus simple de ces formes : une liste de mots associés à des catégories sémantiques indiquant s'ils font référence à une personne, à un lieu ou à une autre entité »

Les méthodes à base de dictionnaire, l'une des approches les plus fondamentales et les plus intuitives, s'appuyer sur des listes complètes d'entités pour identifier les occurrences d'entité dans le texte [Simpson and Demner-Fushman, 2012]. Ces systèmes déterminent si un mot ou un groupe de mots sélectionné dans le texte correspond exactement à un terme issu d'une ressource lexicale. Lorsqu'elles sont utilisées comme méthodes autonomes, les approches à base de dictionnaire font généralement preuve d'une précision raisonnablement élevée, mais leur rappel est médiocre en raison de l'existence d'erreurs d'orthographe et de variantes morphologiques non couvertes par les dictionnaires [Tuason et al., 2004, Ben Abacha and Zweigenbaum, 2011]. Cependant, une faible précision est également possible en raison de l'homonymie et de la polysémie [Ben Abacha and Zweigenbaum, 2011].

Pour ces raisons, certains algorithmes de correspondance de chaîne inexacts sont couramment utilisés pour améliorer la précision et le rappel des approches à base de dictionnaire. Certaines méthodes améliorent les performances en générant d'abord des variantes orthographiques pour les termes d'une ressource biomédicale, puis en ajoutant ces termes supplémentaires aux listes de mots sous-jacentes [Tsuruoka and Tsujii, 2003]. Les méthodes sont ensuite en mesure d'effectuer une correspondance exacte à l'aide de la ressource augmentée. D'autres méthodes effectuent une recherche approximative de chaînes au lieu d'une correspondance exacte [Wang et al., 2009]. Malgré ces améliorations, les méthodes basées sur les dictionnaires sont le plus souvent utilisées conjointement avec des approches de REN plus avancées.

1.3.3.2 À base de règles

Les systèmes à base de règles exigent des connaissances approfondies pour développer des règles d'extraction. Certains systèmes définissent des règles décrivant les patrons de composition des entités nommées et leur contexte, qui sont basés sur la théorie des automates, puis sur les grammaires [Nouvel et al., 2016]. D'autres systèmes utilisent des règles basées sur des patrons qui exploitent les caractéristiques orthographiques et lexicales de types d'entités ciblées, par exemple pour reconnaître les noms de protéines [Fukuda et al., 1998].

Un grand avantage des systèmes à base de règles est qu'ils n'ont pas besoin d'un ensemble de données d'entraînement annotées, ils constituent donc la meilleure option en l'absence de données annotées [Nadeau and Sekine, 2007]. Alors qu'ils obtiennent généralement de meilleures performances que les approches à base de dictionnaire [Simpson and Demner-Fushman, 2012], la génération manuelle des règles

requises est un processus coûteux et fastidieux, et, comme elles sont généralement très spécifiques pour atteindre une précision élevée, elles sont difficiles pour étendre à d'autres types d'entité [Simpson and Demner-Fushman, 2012].

1.3.3.3 À base d'apprentissage automatique

L'accès facile à la puissance de calcul stimule le développement des systèmes de REN basés sur des techniques d'apprentissage automatique. Les différentes approches statistiques peuvent être supervisées, non supervisées ou semi-supervisées [Nadeau and Sekine, 2007].

Supervisé Les systèmes utilisant des approches supervisées considèrent la reconnaissance d'entité nommée comme une tâche de catégorisation ou d'étiquetage de séquence à plusieurs classes.

Les classifieurs couramment utilisés pour la REN comprennent les classifieurs *Naïve Bayes* [Rish, 2001] et *Support Vector Machine* (*SVM*) [Cortes and Vapnik, 1995, Mitsumori et al., 2005]. La performance globale des approches basées sur la catégorisation dépend en grande partie du choix des caractéristiques (*features* en anglais) utilisées pendant l'entraînement [Simpson and Demner-Fushman, 2012].

Au contraire des approches basées sur la catégorisation, les systèmes REN basés sur les séquences prennent en compte des séquences complètes de mots au lieu de mots ou de phrases uniques. Un cadre statistique commun utilisé pour le REN du domaine biomédical est le modèle de Markov caché (*Hidden Markov Model*, *HMM*) [Eddy, 1996, Ekbal and Bandyopadhyay, 2007, Morwal et al., 2012]. Les méthodes basées sur le modèle de Markov d'entropie maximale (*MEMM*) sont également courantes [Fresko et al., 2005, Finkel et al., 2004]. Cependant, il est souvent démontré que les champs aléatoires conditionnels (*Conditional Random Fields*, *CRF*) [Lafferty et al., 2001] constituent des algorithmes statistiques supérieurs pour la REN du domaine biomédical [Ben Abacha et al., 2015b, Leaman et al., 2015].

Toutes les méthodes d'apprentissage supervisé pour la REN, fondées sur la catégorisation ou sur l'étiquetage des séquences, reposent sur un grand volume de données d'entraînement annotées à partir desquelles le modèle infère des connaissances. Les éléments clés d'une performance élevée sont la disponibilité d'un vaste ensemble de données d'entraînement étiquetées et d'un ensemble de caractéristiques soigneusement sélectionnées [Ben Abacha and Zweigenbaum, 2011].

Non-supervisé Contrairement à l'apprentissage supervisé, des techniques d'apprentissage non-supervisé sont appliquées aux données non étiquetées et s'appuient sur des structures lexicales et des statistiques. Les données non étiquetées sont généralement faciles à collecter, mais elles sont moins utiles et peuvent souvent nuire à la performance [Carlson et al., 2010].

Le *clustering* est la méthode non supervisée la plus courante pour la REN [Nadeau and Sekine, 2007]. En gros, *clustering* signifie regrouper les points de données non-annotées de manière appropriée [Jain et al., 1999]. L'un des algorithmes les plus connu pour *clustering* est l'algorithme *k-means* [McQueen, 1967].

Semi-supervisé L'apprentissage semi-supervisé tente de résoudre le problème de la collecte d'une grande quantité de données étiquetées en utilisant seulement un petit nombre de données annotées combinées à une plus grande quantité de données non-étiquetées. Cette approche est prometteuse et peut s'avérer très utile si le nombre de données étiquetées est insuffisant [Liao and Veeramachaneni, 2009, Carlson et al., 2010, Zafarian et al., 2015]. Cependant, la plupart des systèmes n'ont pas atteint une précision semblable à celle des approches supervisées [Carlson et al., 2010].

1.3.3.4 Hybrides

Les approches hybrides combinent plusieurs types d'approches pour exploiter les avantages et éviter les limitations de chaque type d'approches [Ben Abacha and Zweigenbaum, 2011]. En général, une étape de post-traitement est nécessaire pour traiter les résultats contradictoires de plusieurs approches. Les approches hybrides produisent généralement de meilleurs résultats que chaque composant [Ben Abacha and Zweigenbaum, 2011, Simpson and Demner-Fushman, 2012, Liu et al., 2015].

1.3.4 Mesures d'évaluation

Les mesures d'évaluation habituelles dans le domaine de fouille de textes sont rappel (Formule 1.1), précision (Formule 1.2) et F-mesure (Formule 1.3, avec $beta = 1$) Les performances des systèmes évaluées avec les mesures habituelles dans le domaine de l'extraction d'information :

Le rappel (Formule 1.1) mesure le nombre d'éléments correctement étiquetés par le système (vrais positifs) rapporté au nombre d'éléments étiquetés dans la référence (vrais positifs et faux négatifs).

$$\text{Rappel} = \frac{\text{vrais positifs}}{\text{vrais positifs + faux négatifs}} \tag{1.1}$$

La précision (Formule 1.2) mesure le nombre d'éléments correctement étiquetés par le système (vrais positifs) rapporté au nombre total d'éléments étiquetés par le système (vrais et faux positifs).

$$\text{Précision} = \frac{\text{vrais positifs}}{\text{vrais positifs + faux positifs}} \tag{1.2}$$

La F-mesure (Formule 1.3) est la moyenne harmonique pondérée du rappel et de la précision. La valeur accordée à β permet de pondérer le rappel ou la précision, ou d'équilibrer les deux mesures (avec $\beta = 1$).

$$\text{F-mesure} = \frac{(1+\beta^2) \times \text{ précision } \times \text{ rappel}}{\beta^2 \times \text{ précision } + \text{ rappel}} \tag{1.3}$$

1.3.5 Schéma d'annotation : BIO

Pour les méthodes basées sur l'apprentissage automatique supervisé, chaque token est annoté avec une étiquette de type indiquant si le token fait partie d'une entité et sa position dans cette entité. `BIO` [Ramshaw and Marcus, 1995] est l'un des schémas d'annotation les plus utilisés [Simpson and Demner-Fushman, 2012], qui

dicte que les tokens individuels classés comme étant soit au début (**B**-egin) d'une entité, à l'intérieur (**I**-inside) des frontières d'une entité, ou en dehors (**O**-utside) des frontières d'une entité.

Phrase : Le brocoli contient une quantité élevée de vitamine K.	
BIO	
Le	O
brocoli	**B-food**
contient	O
une	O
quantité	O
élevée	O
de	O
vitamine	**B-component**
K	**I-component**
.	O

TABLE 1.1 – Un exemple d'annotation au schéma *BIO* d'une phrase extraite de notre corpus.

Le tableau 1.1 illustre un exemple d'annotation au schéma BIO d'une phrase extraite de notre corpus. De plus, il existe également d'autres schéma d'annotation, tel que BILOU [Ratinov and Roth, 2009], Tie or Break [Shang et al., 2018], etc., qui ont chacun ses avantages et ses limites et ne seront pas détaillés dans ce travail.

1.3.6 Ressources

Cette section est consacrée aux ressources du domaine qui sont disponibles en français.

Aliments et aliments complémentaires

- MeSH bilingue anglais-français
  - Le MeSH [6] (*Medical Subject Headings*) est le thésaurus de référence dans le domaine biomédical. La NLM (U.S. National Library of Medicine), qui l'a construit et le met à jour chaque année, l'utilise pour indexer et permettre d'interroger ses bases de données, notamment MEDLINE/PubMed. La traduction en français est réalisée par l'Inserm.
  - La MeSH contient un nœud porte sur des aliments, qui sera interessant pour notre travail.
- OpenFoodFacts
  - La base est riche en termes de contenu, mais elle est très bruitée à cause des erreurs de OCR.
- Table de composition nutritionnelle des aliments Ciqual 2017 [7](France)
- Base de données Oqali [8](France)
- Base de suisse des valeurs nutritives [9](Suisse)

6. http://mesh.inserm.fr/FrenchMesh/
7. https://ciqual.anses.fr/#/cms/telechargement/node/20
8. https://www.oqali.fr/Donnees-publiques
9. https://www.valeursnutritives.ch/fr/telechargement/

— Données nutritionnelles [10](Canada)
— Liste des compléments alimentaires déclarés en France [11]

Médicaments

— Base de données publique des médicaments [12]
  — la base contient uniquement des médicaments avec autorisation active et commercialisés dans le marché en France)
— Répertoire des Spécialités Pharmaceutiques [13]
  — *N.B.* : les deux ressources sont issues de l'ANSM, il suffit d'en choisir une.
— ROMEDI :Référentiel Ouvert du Médicament
  — Cette ontologie est développée à base de BDMP, en ajoutant des relations liées à d'autres ressources ontologiques en anglais.
— ATCFRE :Classification ATC (anatomique, thérapeutique et chimique) [14]

1.4 Conclusion

La reconnaissance des entités biomédicales vise à extraire des informations en identifiant et en catégorisant les termes biomédicaux dans des textes. Dans ce chapitre, nous avons passé en revue des travaux existants, qui permettent de traiter la problématique de la reconnaissance d'entités nommées dans le domaine biomédical. Nous avons présenté des ressources qui sont appliquées aux données biomédicales et qui ciblent les cas d'usages biomédicaux, en mettant l'accent sur les médicaments/aliments. Les nouvelles connaissances tirées de cette recherche sur des travaux connexes aideront à prendre des décisions éclairées en matière de conception de notre système.

10. https://www.canada.ca/fr/sante-canada/services/aliments-nutrition/saine-alimentation/donnees-nutritionnelles/fichier-canadien-elements-nutritifs-fcen-2015.html
11. https://data.economie.gouv.fr/explore/dataset/liste-des-complements-alimentaires-declares/information/
12. Source : http://base-donnees-publique.medicaments.gouv.fr/telechargement.php (mise-à-jour :25/06/2019)
13. http://agence-prd.ansm.sante.fr/php/ecodex/telecharger/telecharger.php (mise-à-jour :25/06/2019)
14. http://bioportal.lirmm.fr/ontologies/ATCFRE

CHAPITRE 2

MATÉRIEL ET MÉTHODES

Sommaire

2.1 Introduction

Après une étude systématique des méthodes et des ressources de l'état de l'art en matière de reconnaissance des entités biomédicales, nous pouvons sans risque conclure à la pénurie de recherches sur la reconnaissance des entités aliments/médicaments à partir de textes biomédicaux pour la langue française. Par conséquent, l'objectif principal de ce travail est d'extraire automatiquement les entités en question à partir de données non étiquetées.

Ce travail est ensuite axé sur deux tâches principales : créer un corpus de test en français dédié aux interactions entre les aliments et les médicaments avec une intervention humaine minimale ; proposer des méthodes qui permettent la reconnaissances automatiquement des entités liées aux aliments/médicaments.

Ce chapitre décrit la méthodologie élémentaire de ce travail. Dans la Sous-section 2.2, les méthodes de construction d'un corpus d'interactions aliments-médicaments en français sont détaillées. Ensuite, dans les Sous-sections 2.3 et 2.4, les méthodes choisies pour la reconnaissance des entités aliments/médicaments sont expliquées.

Après une brève description des outils, des bibliothèques et des ressources nécessaires à la mise en œuvre, le chapitre se termine par un bref résumé des informations présentées.

2.2 Constitution du corpus

En raison du manque de corpus annotés disponibles concernant les interactions aliments-médicaments pour la langue française et afin d'évaluer les méthodes de reconnaissance des entités aliments/médicaments proposées dans ce travail, nous avons créé un corpus de test. Dans cette sous-section, nous présenterons nos méthodes relatives à la constitution d'un corpus d'interactions aliments-médicaments en français, de la sélection à la collection, à la catégorisation, au filtrage, en passant par les modules de nettoyage et de normalisation. A l'issue de cette chaîne de traitements, la performance des méthodes est évaluée.

2.2.1 Sélection de source

Comme mentionné dans la Section 1.2, la littérature scientifique est le genre dominant dans les tâches de fouille de textes du domaine biomédical. L'intuition est de constituer un corpus à partir des publications scientifiques en langue française sur les interactions aliments-médicaments, selon des critères similaires décrites dans [Hamon et al., 2017] pour la construction du corpus POMELO. Une requête avec « interactions aliments-médicaments[1] », le descripteur du thésaurus MeSH de Medline/PubMed dans sa traduction française a été effectué sur CISMeF[2], Catalogue et Index des Sites Médicaux de langue Française. Cette requête, même relâchée, n'a envoyé que 16 résultats, un volume loin d'être satisfaisant pour mener des expériences pour les tâches de la fouille de textes dans le domaine biomédical.

Cette observation nous conduit à utiliser des fiches encyclopédiques, rédigés et/ou révisés par des experts, provenant des sites spécialisés du domaine de la santé, notamment Doctissimo[3] et Passeport Santé[4]. Les deux sites, Doctissimo et Passeport Santé, ont chacun un portail encyclopédique consacré aux produits de santé naturel/compléments alimentaires, tels que ail, ananas, caféine, pamplemousse, réglisse, etc. Chaque fiche présente de manière exhaustive un produit de santé naturel, couvrant la description botanique, l'historique d'utilisation, les propriétés médicinale, la composition, l'utilisation, les recherches, les indications, les précautions, les interactions, l'avis du médecin, les références, etc.

2.2.2 Collection, catégorisation et filtrage d'information

Une chaîne de traitements qui permet la collection, la catégorisation, le filtrage ainsi que le nettoyage et la normalisation est proposée ici.

Tous les documents collectés ne sont pas pertinents par rapport aux besoins. Par exemple, [Alphonse et al., 2006] montre que seul 3% des résumés de la base de don-

1. Identifiant d'origine : D018565 ; CUI UMLS : C0242785
2. http://www.chu-rouen.fr/cismef/
3. https://www.doctissimo.fr/html/sante/phytotherapie/plante-medicinale/guide-phyto.htm
4. https://www.passeportsante.net/fr/Solutions/PlantesSupplements/Index.aspx

nées bibliographique Medline sont pertinents pour identifier les relations d'interactions entre gènes et protéines. Il convient de passer une étape de « trie », c'est-à-dire de sélectionner les documents pertinents et de filtrer davantage des zones « utiles », car le filtrage permet d'éliminer toutes les portions de texte inutiles pour mettre l'accent sur les zones les plus pertinentes pour la tâche [Nazarenko, 2005]. D'après [Balvet, 2002], le filtrage consiste à sélectionner des documents tirés d'un flux d'information textuelle, sur la base d'une comparaison binaire (correspondance/-non correspondance) entre le profil informatif de chaque document et celui du besoin en information. Ici, cette tâche est en deux grandes étapes : catégorisation binaire qui permet de récupérer des documents pertinents ; filtrage qui permet de filtrer davantage dans ces documents des zones pertinents, en l'occurrence des paragraphes contenant des messages sur des interactions aliments-médicaments.

2.2.2.1 Collection

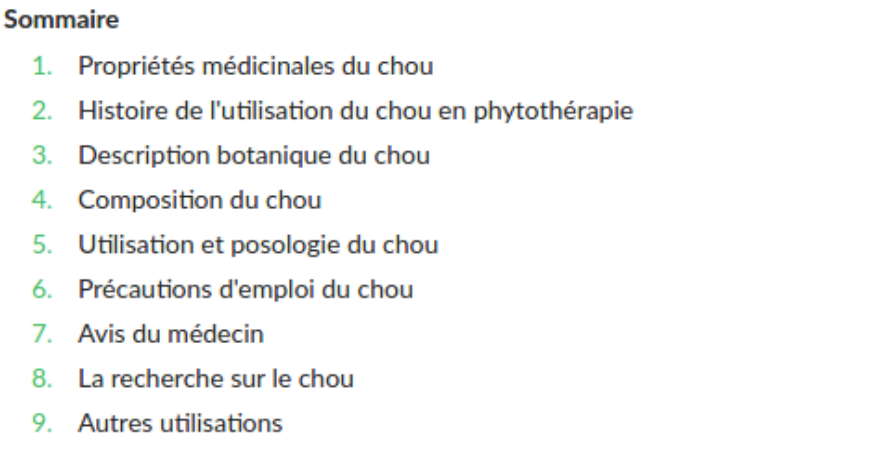
À partir de deux URLs de base (adresses de portail encyclopédique), un fichier json composé d'une liste d'URLs candidats $\langle clé : nom\_fichier, valeur : URL \rangle$ est généré et dédoublonnée.

2.2.2.2 Catégorisation à base de règles

Il s'agit de passer l'objet de page à une suite de règles (méta-données, mots amorces) pour aboutir à une catégorisation binaire : POS (document qui contient des messages concernant des interactions aliments-médicaments) et NEG (document qui n'en contient aucun). Les textes POS et NEG sont stockés séparément.

2.2.2.3 Filtrage à base de règles

Le filtrage est en deux étapes : localisation et extraction. Il s'agit de détecter les frontières des passages contenant des informations sur des interactions aliments-médicaments pour ensuite extraire ces passages pertinents.


Sommaire

1. Propriétés médicinales du chou
2. Histoire de l'utilisation du chou en phytothérapie
3. Description botanique du chou
4. Composition du chou
5. Utilisation et posologie du chou
6. Précautions d'emploi du chou
7. Avis du médecin
8. La recherche sur le chou
9. Autres utilisations

FIGURE 2.1 – Structure possible d'une fiche encyclopédique, exemple extrait du Doctissimo

La figure 2.1 extrait du `Doctissimo` montre une structure possible d'une fiche encyclopédique. A savoir, toutes les sections ne sont pas forcément présentes dans une fiche. Il existe des documents qui n'ont pas une telle section : « interaction » par exemple. Cependant, la classe et les frontières du contenu d'interactions ne sont pas ambiguës, elles sont clairement marquées par des *headings*. Ainsi, une suite de règles de localisation sont créée et employée :

— méta-données : balise *heading* de sections, id/classe `CSS` ;
— présence de lignes blanches avant/après ($fenêtre \pm 1$) ;
— longueur de la ligne en nombre de tokens ($fenêtre \pm 1$, avec un seuil de 15) ;
— expressions régulières avec des mots clés.

Les passages pertinents sont ensuite extraits et stockés pour pouvoir tester de différentes approches d'extraction des entités aliments/médicaments dans la suite de notre travail.

2.2.2.4 Nettoyage et normalisation

Le nettoyage et la normalisation des données est effectué avec le remplacement de chaîne de caractères à l'aide d'expressions régulières. Les cas suivants sont considérés dans cet ordre lors du pré-traitement.

— **Encodage** : en `UTF-8` ;
— **Caractères spéciaux** : par exemple, tabulation, espace insécable, multiples espaces consécutifs $\rightarrow$ un espace blanc ; retours à la ligne $\rightarrow$ '$\backslash n'$;
— **Ligatures** : par exemple, œ $\rightarrow$ oe, æ $\rightarrow$ ae ;
— **Ponctuations** : par exemple, les différentes formes d'apostrophes typographiques, de guillemets.

2.2.3 Résultat et discussion

Afin d'évaluer la performance des méthodes de catégorisation et de filtrage d'information pour s'assurer la validité et la pertinence des résultats obtenus, un corpus de test (appelé supra. `corpus120`) est construit à partir de 120 `URL`s candidats choisis au hasard. Les résultats des deux systèmes sont évalués manuellement avec les mesures d'évaluation en terme de `précision`, `rappel` et `F1-mesure`, présentées dans la Sous-section 1.3.4. Pour le filtrage d'information, il s'agit d'une évaluation stricte, en l'occurrence, la variation de frontière n'est pas autorisée.

Catégorisation Le tableau 2.1 montre les résultats de la catégorisation binaire, seul la catégorie `POS` (documents qui sont pertinents aux interactions aliments-médicaments) est prise en compte lors de calcul.

Catégorisation	VT	Rappel	Précision	F1-mesure
Document(POS)	72	98,63%	97,30%	97,96%

TABLE 2.1 – Résultats du système de catégorisation binaire appliqué sur le corpus de test (corpus120)

Filtrage Le tableau 2.2 illustre les résultats du système de filtrage. L'évaluation du système est stricte, c'est-à-dire que les erreurs de frontières ne sont tolérées, seul

les passages correctement délimités et extraits sont pris en compte.

Filtrage	VT	Rappel	Précision	F1-mesure
Passage(POS)	67	91,78%	97,10%	94,37%

TABLE 2.2 – Résultats du système de filtrage appliqué sur le corpus de test (corpus120)

Les deux systèmes, à base de règles, obtiennent une performance satisfaisante en terme de précision, rappel et F1-mesure. La performance de la catégorisation au niveau du document surpasse celle du filtrage d'information au niveau du passage en terme de rappel, 98,63% et 91,78% respectivement. Une analyse des erreurs montre que les frontières, notamment la fin de passages pertinents, sont mal identifiées à cause de la couverture des règles générées. Néanmoins, ces deux systèmes sont applicables à l'ensemble des URLs candidats.

2.2.4 Statistique du corpus

Après l'application des deux systèmes à l'ensemble des 583 URLs candidats et une correction manuelle en aval, nous avons pu obtenir 292 documents contenant uniquement des passages pertinents. Le tableau 2.3 présente des statistiques descriptives globales du corpus.

	Doctissimo		Passeport Santé	
Catégorie/Pertinence	POS	NEG	POS	NEG
Nb de documents	62	188	230	103
Nb de tokens	65 358	125 339	660 193	177 683
Nb moyens de tokens par document	1 054	667	2 870	1 725
Nb de tokens (passages extraits)	1 660	–	16 344	–
Nb moyens de tokens par passages	27	–	71	–
Différence de tokens (%)	-97,4%	–	-97,5%	–

TABLE 2.3 – Statistiques descriptives du corpus d'interactions aliments-médicaments en français

Il est à noter que le filtrage a significativement réduit la taille du corpus, soit une réduction de taille d'environ 97,4% en terme de nombre de tokens. Ce résultat correspond à l'observation dans [Alphonse et al., 2006], présentée dans 2.2.2. Cela signifie que nous pouvons nous appuyer sur cette méthode sur l'ensemble de documents pour concentrer uniquement sur les passages mentionnant des interactions.

2.3 REN aliments/médicaments à base de dictionnaire

Les approches basées sur un dictionnaire (lexique, gazetier, liste de mots, ...) identifient les noms d'entités en faisant correspondre les termes dans un dictionnaire à des textes en entrée. Inspirés des travaux connexes présentés *supra.* dans la Sous-section 1.3.3.1, nous adoptons deux approches : recherche exacte et recherche approximative.

2.3.1 Recherche exacte

2.3.1.1 Algorithmes

Pour la recherche exacte, l'algorithme le plus facile à implémenter est la recherche à force brute. Mais il est coûteux en terme de complexité du temps. Soit m la taille du dictionnaire, n la taille du texte à rechercher, la complexité temporelle pour une recherche à force brute est $\Theta(nm)$. Cela n'est pas efficace pour un système ayant un grand volume de données.

Nous profitons donc l'algorithme de recherche `Aho-Corasick` (`automate AC`) en combinant la structure de données `Trie` pour le dictionnaire. La complexité de temps ne dépend pas du nombre de mots-clés présents dans le dictionnaire (m), elle sera de $\Theta(n)$.

2.3.1.2 Chaîne de traitements

Le schéma 2.2 illustre la chaîne de traitements à base de dictionnaire par recherche exacte. La première étape consiste à construire un dictionnaire du domaine à partir des ressources disponibles en Trie. Le système prend en entrée un corpus de textes bruits, segmente les textes en phrases, ensuite effectue la recherche des entités, et produit un fichier pivot au format `BRAT` pour des traitements en aval.

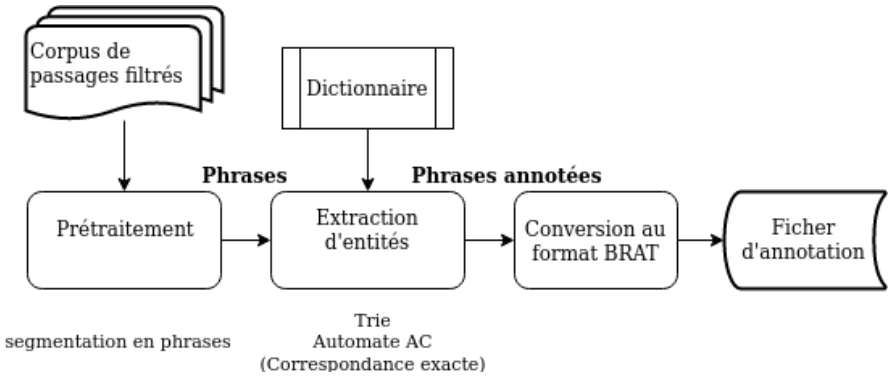


FIGURE 2.2 – Chaîne de traitements à base de dictionnaire par recherche exacte

La performance de ce type de système dépend non seulement la couverture de dictionnaire, mais aussi la qualité de texte.il atteint généralement une bonne précision, mais souffre d'un faible rappel. En effet, il existe des fautes d'orthographe ou des variantes de noms d'entité non couverts par les dictionnaires. Par conséquent, la correspondance approximative est utilisée pour but d'améliorer le rappel.

2.3.2 Recherche approximative

Pour la recherche approximative, nous utilisons la distance d'édition, en l'occurrence la distance de Levenshtein en misant le seuil de confidence à 75% (ratio normalisé par rapport à la taille de patron).

Le schéma 2.3 illustre la chaîne de traitements par recherche approximative. La recherche est effectuée uniquement sur les portions de textes qui n'ont pas préalablement été annotées par la recherche exacte. Cela est efficace et réduit le taux de bruits.

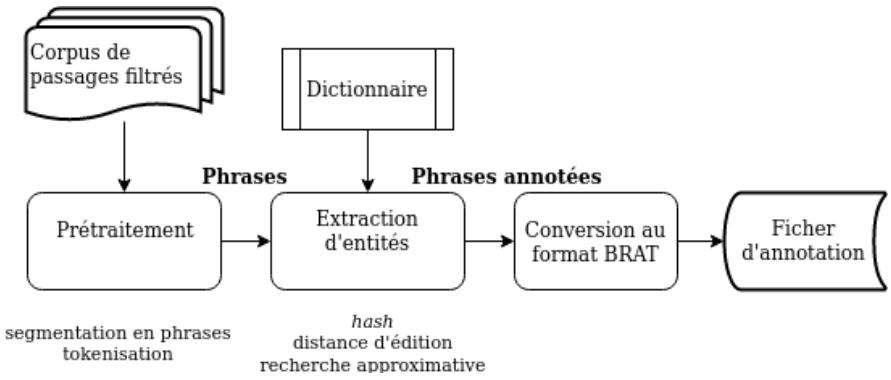


FIGURE 2.3 – Chaîne de traitements à base de dictionnaire par recherche approximative

2.4 REN aliments/médicaments par apprentissage automatique

Comme présentés dans 1.3.3.3, `CRF` constituent des algorithmes statistiques supérieurs pour la REN du domaine biomédical. Dans ce travail, nous reposons uniquement sur les `CRF` pour l'extraction d'entités aliments/médicaments par apprentissage automatique. Nous présentons d'abord nos choix de caractéristiques, ensuite les configurations de l'outil (en l'occurrence `sklearn-crfsuite`), notamment les hyperparamètres : algorithme d'optimisation et la régularisation.

2.4.1 Sélection de caractéristiques

Les caractéristiques utilisés sont des caractéristiques morphologiques, morpho-syntaxiques et sémantiques.

— **caractéristiques de surface : morphologiques** :
  * token ; fenêtre $\pm$ 1 ;
  * bigramme de tokens ; fenêtre $\pm$ 1 ;
  * casse typographique de token
    — TOUS EN MAJUSCULES ;
    — Première lettre en majuscule ;
    — tous en minuscules
  * taille de token $taille > 4 \ ? \ true : false$;
  * signe de ponctuation $signedeponctuation \ ? \ true : false$;
  * chiffre $présencedechiffre \ ? \ true : false$;
  * préfixe $token[\ : 4]$;
  * suffixe $token[-4 :]$;

— **caractéristiques de surface :phonétiques** :
  * SOUNDEX[5] ; fenêtre $\pm$ 1 ;
  * bigramme de SOUNDEX ; fenêtre $\pm$ 1 ;

— **caractéristiques morpho-syntaxiques** :

5. *N.B.* : Strictement dit, la caractéristique SOUNDEX n'est pas au niveau morphologique, elle est par contre l'application de l'algorithme phonétique.

    * POS ; fenêtre $\pm$ 1 ;
    * bigrame de POS ; fenêtre $\pm$ 1 ;
    * POS $pos \in \{NC, NP, V, A, ADV, NUM\}$? $true$: $false$; fenêtre $\pm$1 ;

— **caractéristiques sémantiques** :
    * groupe sémantique du token (eg, `food`, `drug`, même dictionnaire utilisé dans 2.3) ; fenêtre $\pm$ 1.

2.4.2 Configuration d'hyper-paramères

Algorithme d'optimisations

— `lbfgs` - *Gradient descent* utilisant la méthode L-BFGS ;
— `l2sgd` - *Stochastic Gradient Descent* avec la $l2$ regularization ;
— `ap` - *Averaged Perceptron* ;
— `pa` - *Passive Aggressive (PA)* ;
— `arow` - *Adaptive Regularization Of Weight Vector (AROW)* ;

Ici, nous choisissons `lbfgs` (*Limited-memory Broyden-Fletcher-Goldfarb-Shanno*), et `sklearn-crfsuite` se chargera du reste.

Régularisation

`Elastic Net` qui utilise en même temps $l1$ et $l2$.

2.5 Matériel

2.5.1 Lexique

Les ressources linguistiques, ontologiques, terminologiques, thésaurus et données ouvertes liées en langue française utilisés pour la construction de dictionnaires. Les ressources sont présentées *supra.* dans la Sous-section 1.3.6.

— Table de composition nutritionnelle des aliments `Ciqual` 2017 ;
— Base de données `Oqali` ;
— `ROMEDI` (Référentiel ouvert du Médicament) ;
— `BDPM` (base de données publique des médicaments) ;
— `ATCFRE` (classification `ATC` – anatomique, thérapeutique et chimique) ;
— `MeSH` (Medical Subject Headings) bilingue anglais-français (ressource dans l'`UMLs`) ;
— Liste des compléments alimentaires déclarés ;
— *stopwords* pour le français ;

2.5.2 Outils

— **TAL** : NLTK [6], TreeTagger [7](Python-wrapper [8]), flashtext [9], fuzzywuzzy [10], jel-

6. https://www.nltk.org/
7. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
8. https://perso.limsi.fr/pointal/dev:treetaggerwrapper
9. https://github.com/vi3k6i5/flashtext
10. https://github.com/seatgeek/fuzzywuzzy

lyfish [11](pour distance de Levenshtein et SOUNDEX) ;
— **Apprentissage automatique** : sklearn-crfsuite [12] ;
— **Évaluation** : brateval.jar [13], conlleval [14].
— **Divers** : pandas [15]

2.6 Conclusion

Dans ce Chapitre, nous avons réalisé deux tâches principales : créer un corpus de test en français dédié aux interactions aliments-médicaments avec une intervention humaine minimale ; et proposer des méthodes qui permettent la reconnaissances automatiquement des entités liées aux aliments/médicaments.

La méthodologie de la constitution du corpus a été élaborée au forme d'une chaîne de traitements. Les approches pour extraire les entités aliments/médicaments, à base de dictionnaire et par apprentissage automatique, ont été expliquées. Les ressources et les outils sont également présentées. En nous appuyant sur le matériel et les méthodes proposées, nous préparerons des expériences dans le Chapitre 4 qui suit.

11. https://github.com/jamesturk/jellyfish
12. https://github.com/TeamHG-Memex/sklearn-crfsuite
13. https://bitbucket.org/nicta_biomed/brateval/src/master/
14. https://github.com/sighsmile/conlleval
15. https://pandas.pydata.org/

Deuxième partie

Expérimentations

CHAPITRE 3

EXPÉRIENCES

Sommaire

3.1 Introduction

Dans ce chapitre, le but de notre travail est de concevoir les expériences, de mettre en œuvre et de comparer les approches de reconnaissance d'entités aliments/médicaments proposées au Chapitre 2.

3.2 Corpus

3.2.1 Répartition du corpus

	Train	*Test*
Nb de documents	60	20
Nb de tokens	3 599	1 151
Nb de phrases	192	63
Nb de tokens par phrase	18,7	18,3

TABLE 3.1 – Répartition du corpus

3.2.2 Démarches d'annotation

Nous avons conçu nos démarches en faisant référence aux travaux décrits dans le Guide d'annotation des effets secondaires rapportés par les patients sur les réseaux sociaux [1][Grouin, 2015].

1. https://perso.limsi.fr/grouin/publis/vigi4med-guide-annotation-2015.pdf

Notre guide d'annotation a été appliqué sur les 80 échantillons du corpus par deux annotateurs non-experts [2] du domaine. L'annotation a été réalisée à l'aide de l'outil d'annotation BRAT [Stenetorp et al., 2012]. Les 20 échantillons du corpus de test ont fait objet d'une double annotation, puis d'un consensus entre les deux annotateurs. Les taux d'accord inter-annotateurs ont été calculés sur les 20 échantillons, et évalués en terme de F1-mesure à l'aide de l'outil BRATeval.

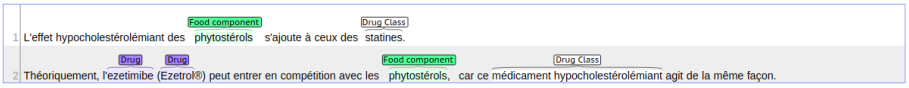


FIGURE 3.1 – Exemple d'un extrait du corpus annoté

L'annotation manuelle a été effectuée à partir d'une pré-annotation par correspondance exacte des mots-clés présents dans le dictionnaire décrit dans la Sous-section 2.3. Le tableau 3.2 liste de manière exhaustive les taux d'accord inter-annotateurs entre les résultats de la pré-annotation, des deux annotateurs et du consensus. L'évaluation couvre les 4 types d'entités intéressées.

Annotateurs		Global	Food	Component	Drug	DrugClass
A	Pré-annot.	0,50	0,67	0,11	0,66	0,28
B	Pré-annot.	0,59	0,71	**0,00**	0,70	0,44
A	B	**0,86**	0,93	**0,56**	0,84	**0,87**
A	Consensus	0,98	**1,00**	0,97	0,99	0,96
B	Consensus	0,87	0,93	0,58	0,86	0,88
Consensus	Pré-annot.	**0,57**	**0,64**	**0,00**	0,72	**0,42**

TABLE 3.2 – Taux d'accord inter-annotateurs (F1-mesure) sur les 20 échantillons du corpus de test.

Les deux annotateurs obtiennent des accords élevés (soit F1-Mesure=0,86), qui confirment qu'ils ont compris le guide d'annotation de la même manière. Cependant, la pré-annotation par correspondance exacte fournit un résultat de qualité moyenne (F1-Mesure=0,57 au niveau global, avec une meilleure réussite sur les drug mais aucun apport sur les component).

Une fois l'annotation manuelle sur les 80 échantillons terminée, les fichiers tabulaires sont convertis au schéma BIO, prêts pour passer aux expériences par apprentissage automatique.

3.3 Configurations

3.3.1 Hypothèses

Partant sur les principes que moins les caractéristiques de surface, plus de gain de performance. Nous avons donc formuler les hypothèses suivantes pour étudier les impacts de différents types de caractéristiques sur la performance du système :

2. Voir. [Hernández et al., 2014] pour plus d'informations sur l'annotation par non-experts.

a. De manière générale, l'étiquetage en POS peut améliorer la performance de l'extraction automatique d'entités aliments/médicaments ;
b. Pour un jeu de données ayant une annotation morpho-syntaxique réalisée de manière automatique (donc moins de confiance), ressources externes (en l'occurrence Soundex) peut diminuer les erreurs de tokens et réduire la « taille » de caractéristiques.
c. L'information sur le type sémantique d'un token permet d'améliorer la performance.

3.3.2 Études ablatives

Conformément à nos hypothèses, nous proposons 7 combinaisons possibles des types de caractéristiques indiquées dans la Section 2.4.

Combo.1 baseline : caractéristiques de surface (morphologiques) ;
Combo.2 baseline − tokens + caractéristiques morphosyntaxiques (POS) ;
Combo.3 baseline − tokens + caractéristiques de surface (SOUNDEX) ;
Combo.4 baseline + caractéristiques de sémantiques ;
Combo.5 baseline − tokens + POS + caractéristiques de sémantiques ;
Combo.6 baseline − tokens + caractéristiques morphosyntaxiques (POS) + caractéristiques de surface (SOUNDEX) ;
Combo.7 tous.

3.4 Conclusion

Dans ce chapitre, nous avons décrit les démarches d'annotation du corpus et avons conçu une suite de différentes configurations d'expériences pour l'extraction des entités aliments/médicaments dans le corpus en question.

CHAPITRE 4

RÉSULTATS ET DISCUSSION

Sommaire

4.1 Introduction

Une série d'expériences a été réalisée pour évaluer les approches supervisées présentées. Les systèmes sont évalués sur les 20 échantillons du corpus de test préalablement construits et annotés en entités aliments/médicaments. Les mesures d'évaluation utilisées sont la précision, le rappel et la F1-mesure, présentées dans 1.3.4. Les sections suivantes montrent les résultats de l'étude ablative et discutent de résultats intéressants.

4.2 Résultats et discussion

Configurations	Rappel	Précision	F1-mesure	δ
baseline	**0,22**	0,73	**0,34**	NaN
baseline − tokens + (POS)	0,37	0,71	0,49	0,15
baseline − tokens + (SOUNDEX)	0,41	0,70	0,52	0,18
baseline + caractéristiques de sémantiques	0,31	0,68	0,43	0,09
baseline − tokens + POS + caractéristiques de sémantiques	0,40	**0,66**	0,50	0,16
baseline − tokens + (POS) + (SOUNDEX)	0,48	**0,77**	0,59	0,25
...				
tous	**0,51**	0,73	**0,60**	0,26

TABLE 4.1 – Résultats d'une étude ablative sur le corpus de test d'interactions aliments-médicaments en français

Le tableau 4.1 illustre les résultats d'une étude ablative avec toutes les combinaisons possibles des types de caractéristiques évalués sur le corpus de test d'interactions aliments-médicaments en français. Les caractéristiques POS fonctionnent, mais avec une amélioration limitée : 0,15 de gain en terme de `F1-mesure` par rapport au modèle *baseline* ; tandis que les caractéristiques SOUNDEX marchent encore mieux que les caractéristiques POS. Quant aux les catégories (`type`, comme le défini et précisé dans la définition 1.3 et la Sous-section 1.3.2) sémantiques

(apparition d'un token dans le dictionnaire du domaine), elles n'apportent pas d'amélioration significative au système.

Retraçons les résultats du *baseline* construit à base de dictionnaire par correspondance exacte, présentés dans le tableau 3.2, *i.e.*, le taux d'accord inter-annotateurs entre le `consensus` et la `pré-annotation` a une `F1-mesure` de 0,57 au niveau global, meilleure que celle des modèles entraînés avec des informations sémantiques. Le résultat n'est pas évident, certes, explicable : la méthode à base de dictionnaire récupère une suite de tokens qui correspond à un type, tandis que les approches par apprentissage statistique implémentées ici recherche isolément un seul token dans le dictionnaire, cela pourra introduire des bruits.

4.3 Travaux futurs

Améliorations envisagées : nous avons 3 pistes envisageables :

- **Jeu de données** : En l'absence de volumes suffisants de données pertinentes, le corpus conçu dans ce travail a une taille modeste. Il est donc envisageable d'essayer des méthodes de la recherche d'information pour pouvoir récupérer des informations sur les interactions aliments-médicaments dans le monde francophone ;
- **Ingénierie de caractéristiques** : Il convient de contrôler la qualité des caractéristiques : bon nettoyage, réduction de dimension, usage d'autres types de représentation,etc ;
- **Autres approches** : Une fois un volume suffisant de corpus pertinent soit disponible, même sans annotations, sera nous aider à extraire des entités et des relations intéressantes à l'aide de matériel/méthodes émergentes, tel que les modèles de langage pré-entraînes, l'apprentissage profond, etc.

4.4 Conclusion

Dans ce chapitre, nous avons évalué, discuté et comparé des résultats obtenus par les différentes configurations. Les méthodes par apprentissage statistique, ne permettent un résultat saillant. Nous proposons donc quelques pistes exploitables pour nos travaux futurs.

CONCLUSION GÉNÉRALE

Depuis plusieurs années, l'extraction d'informations sur les entités médicamenteuses et les interactions médicamenteuses fait l'objet de nombreuses recherches, compétitions et campagnes d'évaluation. Cependant, la reconnaissance des entités d'aliments/médicaments dans des textes portant sur les interactions aliments/médicaments ne rarement fait l'objet des recherches dans le domaine de la fouille de texte biomédical. Dans ce contexte-là, le Projet ANR MIAM a été initialisé, et s'appuie pour l'instant sur des ressources en anglais. Dans le monde francophone, il existe bien des ressources utiles, mais qu'elles sont souvent hétérogènes et éparses sur le Web. Il est intéressant de les récupérer, trier et regrouper pour en extraire des connaissances utiles et les fournir aux experts de la santé, aux patients et aux grands publiques à des fins d'éviter de mauvaise utilisation de médicaments.

Dans le cadre de ce mémoire, il s'agit notamment de créer un tel corpus en français et proposer de différents méthodes qui permettent de concevoir un système d'extraction des entités aliments/médicaments, pour préparer des tâches, telles que l'extraction de relations entre ces entités (en l'occurrence, interactions aliments-médicaments) en aval.

Dans ce travail, nous avons tout d'abord encadrer notre recherche dans le contexte de fouille de textes du domaine biomédical, ou bien `BioNLP`, qui partagent un large éventail de méthodes et d'outils de traitement en commun. Ensuite, nous avons défini des concepts/notions essentielles pour ce travail, notamment dans des tâches de la reconnaissance d'entités nommées biomédicales. Après une étude systématique des méthodes et des ressources existantes, nous avons proposé et mis en œuvre nos propres méthodes pour la constitution d'un corpus du domaine en question et l'extraction des entités aliments-médicaments par de différentes approches. En fin, nous avons évalué, discuté et comparé les résultats produits par nos systèmes, et finalement proposer des pistes exploitables pour les travaux futurs.

BIBLIOGRAPHIE

[Alphonse et al., 2006] Alphonse, , Aubin, S., Bessieres, P., Bisson, G., Hamon, T., Lagarrigue, S., Nazarenko, A., Manine, A.-P., Nédellec, C., Vetah, M., Poibeau, T., and Weissenbacher, D. (2006). Event-based information extraction for the biomedical domain: the caderige project. *Joint Workshop on Natural Language Processing in Biomedicine and its applications*. – Cité pages 22 et 25.

[Ananiadou et al., 2005] Ananiadou, S., McNaught, J., and Karamanis, N. (2005). Text mining for biology and biomedicine. *Artech House, London*, 33. – Cité page 12.

[Balvet, 2002] Balvet, A. (2002). *Approches catégoriques et non catégoriques en linguistique des corpus spécialisés,application à un système de filtrage d'information*. Sciences de l'homme et société, Université Paris Nanterre. – Cité page 23.

[Ben Abacha et al., 2015a] Ben Abacha, A., Chowdhury, M. F. M., Karanasiou, A., Mrabet, Y., Lavelli, A., and Zweigenbaum, P. (2015a). Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug-drug interaction extraction and classification. *Journal of biomedical informatics*, 58. – Cité page 5.

[Ben Abacha et al., 2015b] Ben Abacha, A., Chowdhury, M. F. M., Karanasiou, A., Mrabet, Y., Lavelli, A., and Zweigenbaum, P. (2015b). Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug-drug interaction extraction and classification. *Journal of biomedical informatics*, 58. – Cité page 17.

[Ben Abacha and Zweigenbaum, 2011] Ben Abacha, A. and Zweigenbaum, P. (2011). Medical entity recognition: A comparison of semantic and statistical methods. – Cité pages 14, 15, 16, 17 et 18.

[Bista et al., 2006] Bista, D., Palaian, S., Shankar, P. R., Prabhu, M., Paudel, R., and Mishra, P. (2006). Understanding the essentials of drug interactions: A potential need for safe and effective use of drugs. *Kathmandu University medical journal (KUMJ)*, 5:421–30. – Cité page 5.

[Björne et al., 2013] Björne, j., Kaewphan, S., and Salakoski, T. (2013). Uturku: Drug named entity recognition and drug-drug interaction extraction using svm classification and domain knowledge. – Cité page 5.

[Bordea et al., 2018] Bordea, G., Thiessard, F., Hamon, T., and Mougin, F. (2018). Automatic query selection for acquisition and discovery of food-drug interactions. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings*, pages 115–120. – Cité page 6.

[Bushra et al., 2011] Bushra, R., Aslam, N., and Khan, A. (2011). Food-drug interaction. *Oman medical journal*, 26:77–83. – Cité page 5.

[Carlson et al., 2010] Carlson, A., Betteridge, J., Wang, R., Hruschka, E., and Tom, M. (2010). Coupled semi-supervised learning for information extraction. pages 101–110. – Cité pages 17 et 18.

[Chapman and Cohen, 2009] Chapman, W. and Cohen, K. (2009). Current issues in biomedical text mining and natural language processing. *Journal of biomedical informatics*, 42:757–9. – Cité page 12.

[Cohen, 2014] Cohen, K. (2014). Biomedical natural language processing and text mining. In Sarkar, I., editor, *Methods in Biomedical Informatics: A Pragmatic Approach*, pages 141–177. – Cité page 12.

[Cohen, 2010] Cohen, K. B. (2010). Bionlp: Biomedical text mining. In Indurkhya, N. and Damerau, F. J., editors, *Handbook of Natural Language Processing*, chapter 21, pages 605–625. Chapman and Hall, second edition. – Cité page 12.

[Cohen and Hunter, 2008] Cohen, K. B. and Hunter, L. (2008). Getting started in text mining. *PLoS Comput Biol*, 4(1):e20. – Cité page 6.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.*, 20(3):273–297. – Cité page 17.

[Eddy, 1996] Eddy, S. R. (1996). Hidden markov models. *Current opinion in structural biology*, 6(3):361–365. – Cité page 17.

[Ekbal and Bandyopadhyay, 2007] Ekbal, A. and Bandyopadhyay, S. (2007). A hidden markov model based named entity recognition system: Bengali and hindi as case studies. In *International Conference on Pattern Recognition and Machine Intelligence*, pages 545–552. Springer. – Cité page 17.

[Ernst, 2017] Ernst, P. (2017). *Biomedical knowledge base construction from text and its applications in knowledge-based systems*. PhD thesis. – Cité page 13.

[Finkel et al., 2004] Finkel, J., Dingare, S., Nguyen, H., Nissim, M., Manning, C., and Sinclair, G. (2004). Exploiting context for biomedical entity recognition: From syntax to the web. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*, JNLPBA '04, pages 88–91, Stroudsburg, PA, USA. Association for Computational Linguistics. – Cité page 17.

[Frankel, 2003] Frankel, E. (2003). Basic concepts. In BJ McCabe, E. F. and Wolfe, J., editors, *Handbook of food-drug Interactions*, page 2. CRC Press. Frankel EH. (2003). Basic Concepts. In: Handbook of food-drug Interactions, McCabe BJ, Frankel EH., Wolfe JJ (Eds.) pp. 2, CRC Press, Boca Raton, 2003. – Cité page 5.

[Fresko et al., 2005] Fresko, M., Rosenfeld, B., and Feldman, R. (2005). A hybrid approach to ner by memm and manual rules. pages 361–362. – Cité page 17.

[Fukuda et al., 1998] Fukuda, K., Tamura, A., Tsunoda, T., and Takagi, T. (1998). Toward information extraction: Identifying protein names from biological papers. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 98:707–18. – Cité page 16.

[Fung et al., 2017] Fung, K. W., Kapusnik-Uner, J., Cunningham, J., Higby-Baker, S., and Bodenreider, O. (2017). Comparison of three commercial knowledge bases for detection of drug-drug interactions in clinical decision support. *Journal of the American Medical Informatics Association : JAMIA*, 24. – Cité page 5.

[Grishman and Sundheim, 1996] Grishman, R. and Sundheim, B. (1996). Message understanding conference-6: A brief history. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, COLING '96, pages 466–471, Stroudsburg, PA, USA. Association for Computational Linguistics. – Cité page 13.

[Grouin, 2015] Grouin, C. (2015). Guide d'annotation des effets secondaires rapportés par les patients sur les réseaux sociaux. Guide d'annotation 2015-07, LIMSI, Orsay, France. – Cité page 33.

[Hamon et al., 2017] Hamon, T., Tabanou, V., Mougin, F., Grabar, N., and Thiessard, F. (2017). Pomelo: Medline corpus with manually annotated food-drug interactions. In *Proceedings of the Biomedical NLP Workshop associated with RANLP 2017*, pages 73–80. – Cité pages 5, 6 et 22.

[Hearst, 2003] Hearst, M. (2003). What is text mining? – Cité page 11.

[Hernández et al., 2014] Hernández, A., Hochheiser, H., Horn, J., Crowley, R., and Boyce, R. (2014). Testing pre-annotation to help non-experts identify drug-drug interactions mentioned in drug product labeling. – Cité page 34.

[Huang and lu, 2015] Huang, C.-C. and lu, Z. (2015). Community challenges in biomedical text mining over 10 years: Success, failure and the future. *Briefings in bioinformatics*, 17. – Cité page 5.

[Jackson, 2003] Jackson, P. (2003). Natural language processing for online applications : Text retrieval, extraction and categorization. *Computational Linguistics*, 29:510–511. – Cité page 11.

[Jain et al., 1999] Jain, A., Murty, M., and Flynn, P. (1999). Data clustering: A review. *ACM Comput Surv*, 31:264–323. – Cité page 17.

[Karsten and Suominen, 2009] Karsten, H. and Suominen, H. (2009). Mining of clinical and biomedical text and data: Editorial of the special issue. *International journal of medical informatics*, 78:786–7. – Cité page 12.

[Lafferty et al., 2001] Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. – Cité page 17.

[Leaman et al., 2015] Leaman, R., Wei, C.-H., and lu, Z. (2015). Tmchem: A high performance approach for chemical named entity recognition and normalization. *Journal of Cheminformatics*, 7:S3. – Cité page 17.

[Leser and Hakenberg, 2005] Leser, U. and Hakenberg, J. (2005). What makes a gene name? Named entity recognition in the biomedical literature. *Briefings in Bioinformatics*, 6(4):357–369. – Cité page 13.

[Li et al., 2018] Li, J., Sun, A., Han, R., and Li, C. (2018). A survey on deep learning for named entity recognition. – Cité page 14.

[Liao and Veeramachaneni, 2009] Liao, W. and Veeramachaneni, S. (2009). A simple semi-supervised algorithm for named entity recognition. *Proceedings of the NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing*. – Cité page 18.

[Lim et al., 2018] Lim, S., Lee, K., and Kang, J. (2018). Drug drug interaction extraction from the literature using a recursive neural network. *PLOS ONE*, 13:e0190926. – Cité page 5.

[Liu et al., 2015] Liu, S., Tang, B., Chen, Q., and Wang, X. (2015). Drug name recognition: Approaches and resources. *Information*, 6:790–810. – Cité pages 15, 16 et 18.

[McQueen, 1967] McQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Computer and Chemistry*, 4:257–272. – Cité page 17.

[Mitsumori et al., 2005] Mitsumori, T., Fation, S., Murata, M., Doi, K., and Doi, H. (2005). Gene/protein name recognition based on support vector machine using dictionary as features. *BMC bioinformatics*, 6 Suppl 1:S8. – Cité page 17.

[Morwal et al., 2012] Morwal, S., Jahan, N., and Chopra, D. (2012). Named entity recognition using hidden markov model (hmm). *International Journal on Natural Language Computing (IJNLC)*, 1(4):15–23. – Cité page 17.

[Nadeau and Sekine, 2007] Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30:3–26. – Cité pages 16 et 17.

[Nazarenko, 2005] Nazarenko, A. (2005). Sur quelle sémantique reposent les méthodes automatiques d'accès au contenu textuel ? In Condamines, A., editor, *Sémantique et corpus*, Traité IC2, série Cognition et traitement de l'information, pages 211–244. Lavoisier. 32 pages. – Cité page 23.

[Névéol et al., 2014] Névéol, A., Grosjean, J., Darmoni, S. J., and Zweigenbaum, P. (2014). Language resources for French in the biomedical domain. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pages 2146–2151. – Cité page 6.

[Névéol et al., 2015] Névéol, A., Grouin, C., Tannier, X., Hamon, T., Kelly, L., Goeuriot, L., and Zweigenbaum, P. (2015). CLEF eHealth evaluation lab 2015 task 1b: clinical named entity recognition. In *Proc of ShARe/CLEF Evaluation Lab*, Toulouse, France. – Cité page 6.

[Nouvel et al., 2016] Nouvel, D., Ehrmann, M., and Rosset, S. (2016). *Named Entities for Computational Linguistics*. Focus Cognitive Science Series. English Translation of the French book. – Cité pages 13, 14, 15 et 16.

[Névéol, 2018] Névéol, A. (2018). *Traitement Automatique de la Langue Biomédicale*. Habilitation à diriger des recherches, Université Paris Sud. – Cité page 12.

[Névéol et al., 2014] Névéol, A., Grouin, C., Leixa, J., Rosset, S., and Zweigenbaum, P. (2014). The QUAERO French medical corpus: A ressource for medical entity recognition and normalization. In *Proc of BioTextMining Work*, pages 24–30. – Cité page 15.

[Oronoz et al., 2015] Oronoz, M., Gojenola, K., Pérez, A., Ilarraza, A., and Casillas, A. (2015). On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions. *Journal of Biomedical Informatics*, 56. – Cité page 6.

[Palleria et al., 2013] Palleria, C., Di Paolo, A., Giofrè, C., Caglioti, C., Leuzzi, G., Siniscalchi, A., Sarro, G., and Gallelli, L. (2013). Pharmacokinetic drug-drug interaction and their implication in clinical management. *Journal of research in*

medical sciences : the official journal of Isfahan University of Medical Sciences, 18:601–610. – Cité page 5.

[Patel and Beckett, 2016] Patel, R. and Beckett, R. (2016). Evaluation of resources for analyzing drug interactions. *Journal of the Medical Library Association : JMLA*, 104:290–295. – Cité page 5.

[Poibeau, 2003] Poibeau, T. (2003). Extraction automatique d'information : Du texte brut au web sémantique. – Cité page 15.

[Ramshaw and Marcus, 1995] Ramshaw, L. A. and Marcus, M. P. (1995). Text chunking using transformation-based learning. *Third ACL Workshop on Very Large Corpora. MIT*, cmp-lg/9505040. – Cité page 18.

[Randriatsitohaina, 2018] Randriatsitohaina, T. (2018). Extraction d'interactions entre aliment et médicament : Etat de l'art et premiers résultats. In *Rencontres des Jeunes Chercheur-euse-s*, Rennes, France. – Cité page 6.

[Ratinov and Roth, 2009] Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 147–155, Stroudsburg, PA, USA. Association for Computational Linguistics. – Cité page 19.

[Rish, 2001] Rish, I. (2001). An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. IBM New York. – Cité page 17.

[Segura-Bedmar et al., 2013] Segura-Bedmar, I., Martínez, P., and Herrero-Zazo, M. (2013). SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics. – Cité pages 5 et 6.

[Shang et al., 2018] Shang, J., Liu, L., Ren, X., Gu, X., Ren, T., and Han, J. (2018). Learning named entity tagger using domain-specific dictionary. In *EMNLP*. – Cité page 19.

[Simpson and Demner-Fushman, 2012] Simpson, M. S. and Demner-Fushman, D. (2012). Biomedical text mining: A survey of recent progress. In Aggarwal, C. C. and Zhai, C., editors, *Mining Text Data*, chapter 14, pages 465–517. Springer Science+Business Media, LLC. – Cité pages 13, 14, 15, 16, 17 et 18.

[Stenetorp et al., 2012] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France. Association for Computational Linguistics. – Cité page 34.

[Sun et al., 2018] Sun, X., Feng, J., Ma, L., Dong, K., and Du, X. (2018). Deep convolution neural networks for drug-drug interaction extraction. pages 1662–1668. – Cité page 5.

[Sørensen, 2002] Sørensen, J. (2002). Herb–drug, food–drug, nutrient–drug, and drug–drug interactions: Mechanisms involved and their medical implications. *Journal of alternative and complementary medicine (New York, N.Y.)*, 8:293–308. – Cité page 5.

[Tsuruoka and Tsujii, 2003] Tsuruoka, Y. and Tsujii, J. (2003). Probabilistic term variant generator for biomedical terms. pages 167–173. – Cité page 16.

[Tuason et al., 2004] Tuason, O., Chen, L., Liu, H., Blake, J., and Friedman, C. (2004). Biological nomenclatures: A source of lexical knowledge and ambiguity. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 238–49. – Cité page 16.

[Wang et al., 2009] Wang, W., Xiao, C., Lin, X., and Zhang, C. (2009). Efficient approximate entity extraction with edit distance constraints. pages 759–770. – Cité page 16.

[Wang et al., 2017] Wang, Y., Liu, S., Rastegar-Mojarad, M., Wang, L., Shen, F., Liu, F., and Liu, H. (2017). Dependency and amr embeddings for drug-drug interaction extraction from biomedical literature. – Cité page 5.

[Wishart et al., 2006] Wishart, D., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006). Drugbank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34:D668–72. – Cité page 6.

[Xu et al., 2018] Xu, B., Xiufeng, S., Zhao, Z., and Vivian, V. (2018). Leveraging biomedical resources in bi-lstm for drug drug interaction extraction. *IEEE Access*, PP:1–1. – Cité page 5.

[Zafarian et al., 2015] Zafarian, A., Rokni, S.-A., Khadivi, S., and Ghiasifard, S. (2015). Semi-supervised learning for named entity recognition using weakly labeled training data. pages 129–135. – Cité page 18.

[Zhang et al., 2019] Zhang, T., Leng, J., and Liu, Y. (2019). Deep learning for drug–drug interaction extraction from the literature: a review. *Briefings in Bioinformatics*. – Cité page 5.

[Zweigenbaum et al., 2007] Zweigenbaum, P., Demner-Fushman, D., Yu, H., and Cohen, K. B. (2007). Frontiers of biomedical text mining: current progress. *Briefings in bioinformatics*, 8(5):358–375. – Cité pages 6, 11 et 12.