Traitement des données massives de santé : Identification et caractérisation des patients résistants aux traitements d'oncologie.

Walid Zeghdaoui

▶ **To cite this version:**

Walid Zeghdaoui. Traitement des données massives de santé : Identification et caractérisation des patients résistants aux traitements d'oncologie.. Autre [cs.OH]. Université de Lyon, 2022. Français. NNT : 2022LYSE2032 . tel-03968442

HAL Id: tel-03968442
https://theses.hal.science/tel-03968442

Submitted on 1 Feb 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° d'ordre NNT : 2022LYSE2032

THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de

L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 512 Informatique et Mathématiques

Discipline : Informatique

Soutenue publiquement le 8 juillet 2022, par :

Walid ZEGHDAOUI

Traitement des données massives de santé.

Identification et caractérisation des patients résistants aux traitements d'oncologie.

Devant le jury composé de :

Florent MASSEGLIA, Directeur de recherche, INRIA, Président

Lydia BOUDJELOUD-ASSALA, Maîtresse de conférences HDR, Université de Lorraine, Rapporteure

Marie-Christine JAULENT, Directrice de recherche, INSERM, Rapporteure

Lynda TAMINE-LECHANI, Professeure des universités, Université Toulouse 3, Examinatrice

Fadila BENTAYEB, Professeure des universités, Université Lumière Lyon 2, Co-Directrice de thèse

Omar BOUSSAID, Professeur des universités, Université Lumière Lyon 2, Co-Directeur de thèse

Contrat de diffusion

Ce document est diffusé sous le contrat *Creative Commons* « **Paternité – pas d'utilisation commerciale - pas de modification** » : vous êtes libre de le reproduire, de le distribuer et de le communiquer au public à condition d'en mentionner le nom de l'auteur et de ne pas le modifier, le transformer, l'adapter ni l'utiliser à des fins commerciales.

UNIVERSITÉ LUMIÈRE LYON 2

ÉCOLE DOCTORALE N°512

INFORMATIQUE ET MATHÉMATIQUES (InfoMaths)

THÈSE

pour obtenir le titre de

Docteur en Informatique

Présentée et soutenue par

Mohamed Walid ZEGHDAOUI

Traitement des données massives de santé : identification et caractérisation des patients résistants aux traitements d'oncologie

---

Soutenue publiquement le 08/07/2022

Devant le jury composé de :

Pr. Omar Boussaid	Université Lumière Lyon 2	Directeur
Pr. Fadila Bentayeb	Université Lumière Lyon 2	Co-Directrice
Pr. Marie-Christine Jaulent	Directrice de Recherche Inserm	Rapportrice
Pr. Lydia Boudjeloud-Assala	Université de Lorraine	Rapportrice
Pr. Florent Masseglia	Directeur de recherche INRIA	Examinateur
Pr. Lynda Tamine	Université de Toulouse	Examinatrice
M. Frederik Joly	Sword-group	Invité

Remerciements

La thèse est un semblant d'une aventure solitaire, mais elle est en réalité un fruit d'un environnement riche en interactions. J'ai eu la chance à travers cette thèse CIFRE d'être un doctorant à la fois du laboratoire ERIC de l'université Lumière Lyon 2 et de l'entreprise Sword.

Je souhaiterais adresser mes premiers remerciements à ceux qui ont assuré la direction de ces travaux par leurs conseils, leur bienveillance et leur patience : Omar Boussaid, Fadila Bentayeb et Frederik Joly. Omar et Fadila m'ont appris à faire de la recherche, à écrire mes premiers papiers, à présenter mes résultats, et à rédiger. Ils m'ont donné beaucoup de liberté pour poursuivre mes idées et ils m'ont supporté pour les présenter dans de nombreuses conférences. Frederik m'a accompagné au quotidien durant mon travail en entreprise et a su me prodiguer les meilleurs conseils même dans les moments les plus difficiles.

Je voudrais également remercier les rapporteurs de ce document, Marie-Christine Jaulent et Lydia Boudjeloud-Assala, ainsi que les autres membres du jury, Florent Masseglia et Lynda Tamine, pour l'intérêt porté à mes travaux.

Je remercie bien évidemment l'ensemble de l'équipe SID du laboratoire ERIC pour leur bon accueil, en particulier mes camarades : Yassine Ramdane, Abderrazek Azri et Reda Benhissen qui font de ce laboratoire un endroit où il fait bon chercher mais surtout bon vivre.

Je remercie également l'ensemble de mes collègues au sein du département SSL de l'entreprise Sword : Edouard Barthuet, Mathieu Husser, Mathieu Maury, Marion Chalumeau, et tous les autres, pour leur accueil et accompagnement durant toutes ces années, sans qui ces années d'études n'auraient pas eu la même saveur.

Enfin et surtout, je souhaiterais exprimer ma plus grande reconnaissance et gratitude à ma famille, spécialement à mes parents Aziza et Abdelhamid et grands-parents, j'ai de la chance de vous avoir. Merci à mes soeurs Myada et Nour El Houda, et à mon frère Mohamed Amir.

Résumé

Au cours de ces dernières années, l'information au sens large est devenue la pièce maîtresse pour révolutionner les projets de transformation numérique. Encore faut-il savoir l'exploiter d'une manière intelligente pour en tirer tous les bénéfices. L'informatisation des données textuelles concerne plusieurs secteurs d'activité, en particulier le domaine médical. Aujourd'hui, la médecine moderne est devenue presque inconcevable sans l'utilisation des données numériques, qui ont fortement affecté la compréhension scientifique des maladies. Par ailleurs, ces dernières années, les données médicales sont devenues de plus en plus complexes en raison de leur croissance exponentielle. Cette forte croissance engendre une quantité de données importante qui ne permet pas d'effectuer une lecture humaine complète dans un délai raisonnable. Ainsi, les professionnels de santé reconnaissent l'importance des outils informatiques pour identifier des modèles informatifs ou prédictifs à travers le traitement et l'analyse automatiques des données médicales. Notre thèse s'inscrit dans le cadre du projet ConSoRe, et vise à créer des cohortes de patients résistants aux traitements anticancéreux. L'identification de ces résistances nous permet de mettre en place des modèles de prédiction des éventuels risques qui pourraient apparaître pendant le traitement des patients, et nous facilite l'individualisation et le renforcement de la prévention en fonction du niveau de risque estimé. Cette démarche s'inscrit dans le cadre d'une médecine de précision, permettant de proposer de nouvelles solutions thérapeutiques adaptées à la fois aux caractéristiques de la maladie (cancer) et aux profils des patients identifiés. Pour répondre à ces problématiques, nous présentons dans ce manuscrit nos différentes contributions. Notre première contribution consiste en une approche séquentielle permettant de traiter les différents problèmes liés au pré-traitement et à la préparation des données textuelles. La complexité de ces tâches réside essentiellement dans la qualité et la nature de ces textes, et est liée étroitement aux particularités des comptes rendus médicaux traités. Outre les opérations de linguistiques standards telles que la tokenisation ou la segmentation en phrases, nous présentons un arsenal de techniques assez large pour la préparation et le nettoyage des données. Notre deuxième contribution consiste en une approche de classification automatique des phrases extraites des comptes rendus médicaux. Cette approche est constituée essentiellement de deux étapes. La première consiste à entraîner les vecteurs de mots pour représenter les textes de façon à extraire le plus de caractéristiques possibles. La seconde étape est une classification automatique de phrases selon leurs informations sémantiques. Nous étudions pour cela les différents algorithmes d'apprentissage automatique (classique et profond) qui fournissent les meilleures performances sur nos données, et nous présentons notre meilleur algorithme. Notre troisième et dernière contribution majeure est consacrée à notre approche de modélisation des résistances aux traitements d'oncologie. Pour cela, nous présentons deux modèles de structuration des données. Le premier modèle nous permet de structurer les informations identifiées au niveau de chaque document (ou compte rendu). Le second modèle est quant à lui introduit au niveau patient, et permet à partir des informations extraites dans plusieurs comptes rendus d'un même patient,

reconstruire son parcours néoplasique. Cette structuration permet d'identifier les réponses aux traitements et les toxicités, qui constituent des composants élémentaires pour notre approche de modélisation des résistances aux traitements d'oncologie.

Abstract

Information, in its broadest sense, has become the centerpiece to revolutionize digital transformation projects in recent years. However, it is still necessary to know how to exploit it intelligently in order to reap the full benefits. Textual data computerization concerns several sectors of activity, specially the medical field. Today, modern medicine has become almost unthinkable without the use of digital data, which has a significant impact on disease scientific understanding. Moreover, due to its exponential growth, medical data analysis has become more complex. The high growth rate results in a large amount of data that a human can not analyze in a fair length of time. Thus, health professionals are recognizing the importance of IT tools to identify informative or predictive patterns through the automatic processing and analysis of medical data. Our thesis is part of the ConSoRe project, and aims to create cohorts of patients resistant to anticancer treatments. The identification of these resistances allows us to set up predictive models to detect potential risks that could appear during patient treatment, and facilitates the individualisation and reinforcement of prevention according to the estimated level of risk. This approach is part of precision medicine, allowing us to propose new therapeutic solutions adapted both to the characteristics of the disease (cancer) and to the profiles of the patients identified. To address these issues, we present in this manuscript our different contributions. Our first contribution consists in a sequential approach to deal with the different problems related to the preprocessing and preparation of textual data. The complexity of these tasks lies mainly in the quality and nature of these texts, and is closely related to the particularities of the medical reports processed. In addition to standard linguistic operations such as tokenisation or sentence segmentation, we present a broad arsenal of techniques for data preparation and cleaning. Our second contribution consists of an approach for the automatic classification of sentences extracted from medical reports. This approach consists essentially of two steps. The first step consists of training word vectors to represent the texts in order to extract as many features as possible. The second step is an automatic classification of sentences according to their semantic information. We study the different machine learning algorithms (classical and deep) that provide the best performance on our data, and we present our best algorithm. Our third and last major contribution is devoted to our approach to modelling resistance to oncology treatments. For this, we present two models for structuring the data. The first model allows us to structure the information identified at the level of each document (or report). The second model is introduced at the patient level, and allows us to reconstruct the neoplastic history of a patient from the information extracted from several reports. This structuring makes it possible to identify responses to treatments and toxicities, which are elementary components for our approach to modelling resistance to oncology treatments.

Table des matières

Liste des tableaux

Table des figures

Acronymes

ADICAP Association pour le Développement de l'Informatique en Cytologie et en Anatomie Pathologique. 46, 53

ASCII American Standard Code for Information Interchange. 52

ATC Anatomical Therapeutic Chemical. 47

CBOW Continuous Bag-of-Words. v, 25, 26

CCAM Classification Commune des Actes Médicaux. 46

CIFRE Convention Industrielle de Formation par la REcherche. 2

CIM-10 Classification Internationale des Maladies, version 10. 46, 49, 108, 111

CLB Centre Léon Bérard. 106, 109, 110, 112

CLCC Centres de Lutte Contre le Cancer en France. 3

CNIL Commission Nationale de l'Informatique et des Libertés. 51

CNN Convolutionnel Neural Network. v, 6, 7, 10, 28, 39, 40, 81, 82, 89, 91

ConSoRe Continuum Soin Recherche. v, 1–3, 7, 44, 45, 47, 51, 113

CRB Centre de Ressources Biologiques. 45, 48

CRF Conditional Random Fields. 12

CSV Comma-Separated Values. 45

CTCAE Common Terminology Criteria for Adverse Events. 107, 108, 111

DNN Deep Neural Network. 36

DRG Diagnosis Related Group. 45

DSE Dossiers de Santé Électronique. 44

DT Decision Tree. 7, 10, 28, 32, 81, 87, 89

EPC Enquête Permanente Cancer. 44, 45

ERIC Entrepôts, Représentation et Ingénierie des Connaissances. 2

GRU Gated recurrent unit. v, 37–39

HTML Hypertext Markup Language. 4, 16, 49, 52

IA Intelligence Artificielle. 10

IC Institut Curie. 76

IDF Inverse Document Frequency. 24
IE Information Extraction. 22
IR Information Retrieval. 74
IRM Imagerie par Résonance Magnétique. 42, 46, 48
ISO International Organization for Standardization. 52
IT Information Technology. 1

kNN Les k Plus Proches Voisins. 7, 10, 28, 34, 35, 81, 87, 89

LDA Latent Dirichlet Allocation. 24
LIWC Linguistic Inquiry and Word Count. 15
LSA Latent Semantic Analysis. 24
LSI Latent Semantic Indexing. 29
LSTM Long Short Term Memory. v, 37–39, 81, 82, 84–87, 89, 91

MIABIS Minimum Information About BIobank data Sharing. 47
MS Maladie Stable. 102
MS Word Microsoft Word. 4, 16, 45, 49, 52

NB Naïve Bayes Classification. 10, 14, 28, 30, 87, 89
NER Named Entity Recognition. 19, 22, 60, 62, 64, 69, 104, 105

OCR Optical Character Recognition.. 17, 18

PCA Principal Component Analysis. v, 28, 29, 78
PDF Portable Document Format. 4, 16, 45, 49, 52, 55
PET-scan Tomographie par Emission de Positons couplé à un scanner. 42, 103
PMSI Programme de Médicalisation des Systèmes d'Information. 4, 45, 48, 51
POS Part-of-Speech Tagging. 12, 21
PR Progression. 102

RC Réponse Complète. 102
RGPD Règlement Général sur la Protection des Données. 51
RNN Recurrent Neural Network. 6, 7, 10, 28, 36–38, 81, 82, 85
RP Réponse Partielle. 102

SI Système d'Informations. 51
SID Systèmes d'Information Décisionnels. 2
SNOMED Systemized Nomenclature of Medicine. 46, 53
SNOMED-CT Systemized Nomenclature of Medicine-Clinical Terms. 65, 66, 115
SQL Structured Query Language. 45

SSL Search & Semantics Lyon. 2

SVM Support Vector Machine. v, 7, 10, 14, 28, 33, 34, 81, 87, 89

TALN Traitement Automatique du Langage Naturel. 2, 5, 19, 25, 30, 35, 57, 82, 83, 85

TF Term Frequency. 24

TF-IDF Term Frequency Inverse Document Frequency. 24, 36

URL Uniform Resource Locator. 53

UTF-8 Universal Character Set Transformation Format - 8 bits. 52

XML Extensible Markup Language. 45–47, 52, 65

CHAPITRE 1

INTRODUCTION

1.1 Contexte général

Au cours de ces dernières années, l'information au sens large est devenue la pièce maîtresse pour révolutionner les projets de transformation numérique. Encore faut-il savoir l'exploiter d'une manière intelligente pour en tirer tous les bénéfices. Grâce à la technologie, des supports électroniques puissants ont été développés pour contenir et stocker d'énormes quantités de données (textes, images, vidéos, etc.). Ce processus d'informatisation des données accompagné du développement des technologies de l'information a produit un réel bouleversement dans notre vie quotidienne. À l'ère de l'informatique, l'ordinateur permet de manipuler et de traiter des quantités colossales de données. Comme le révèle le journal « Statista Digital Economy Compass », 33 zettaoctets [1] de données numériques ont été créées dans le monde au cours de l'année 2018 [2]. De plus, 80% des informations dans le monde sont actuellement stockées au format texte. Ces chiffres permettent d'entraîner des enjeux majeurs posés par la collecte, le stockage et la transmission d'informations, ainsi que la capacité de pouvoir effectuer des recherches optimisées.

L'informatisation des données textuelles concerne plusieurs secteurs d'activité, en particulier le domaine médical. Aujourd'hui, la médecine moderne est devenue presque inconcevable sans l'utilisation des données numériques, qui ont fortement affecté la compréhension scientifique des maladies. Ces dernières années, les données médicales sont devenues de plus en plus complexes en raison de leur croissance exponentielle. Cette quantité de données ne permet pas d'effectuer une lecture humaine complète dans un délai raisonnable. Ainsi, les professionnels de santé reconnaissent l'importance des outils informatiques pour identifier des modèles informatifs ou prédictifs à travers le traitement et l'analyse automatiques des données médicales (Romaszewski et al., 2019). Ces données constituent une mine d'informations essentielle et contiennent des connaissances de grande importance, sur les plans économique, politique et sociétal, et doivent être impérativement exploitées pour améliorer le secteur de la santé en général, et aider la pratique médicale à atteindre un haut niveau d'efficacité en optimisant

1. $1Zo = 10^{21}$ octets.
2. https://www.statista.com/study/52194/digital-economy-compass/

notamment le processus de soins de santé. Extraire et comprendre ces informations permet d'orienter les praticiens sur plusieurs axes, notamment pour aider à la décision médicale, par exemple pour identifier le diagnostic le plus approprié pour un patient, mais aussi pour faire avancer la recherche clinique ; car les médecins sont constamment à la recherche de nouvelles façons d'innover, notamment dans la lutte contre le cancer, et *in fine* de réduire le taux d'échec thérapeutique. Cependant, extraire des connaissances, parfois tacites, de données médicales de manière intelligente est un problème complexe. Car même s'il existe des données médicales structurées (résultats d'analyses ou données démographiques des patients, etc.), la majeure partie (plus de 85%) des données médicales n'est pas structurée (rapports de radiologie, consultations avec des médecins, etc.), et est représentée sous forme de textes libres. Par conséquent, les méthodes de traitement automatique du langage naturel (TALN) et les techniques de fouille de textes sont des outils importants pour la modélisation, la structuration et la mise à disposition de ces informations, et pour optimiser l'exploitation des connaissances qu'elles contiennent.

Cette thèse, intitulée «*traitement de données massives de santé : identification et caractérisation de patients résistants aux traitements d'oncologie*», s'inscrit dans le cadre du projet *ConSoRe*[3], et est l'aboutissement d'une collaboration CIFRE[4] entre le laboratoire ERIC de l'université Lumière Lyon 2, dont le domaine d'expertise est axé sur des problématiques liées à la modélisation des grands entrepôts de données complexes, la fouille des données massives et peu structurées et les processus d'aide à la décision, et l'entreprise Sword, au sein du département SSL "Search & Semantics Lyon", spécialisée dans le traitement et la valorisation des données.

1.2 Objectifs et motivations

La médecine de précision, aussi appelée médecine personnalisée, est un nouveau paradigme très prometteur pour l'élaboration d'un diagnostic ou de traitements individualisés pour les patients (Weil, 2018). Cependant, elle n'a été décrite que d'une manière informelle plutôt que par des approches pratiques et rigoureuses et des protocoles statistiques permettant sa mise en œuvre dans les milieux de soins de santé. La médecine personnalisée est considérée comme l'un des enjeux majeurs des soins aux patients. Enjeu particulièrement important en oncologie où le développement de résistances aux traitements sont les prémices d'échappements thérapeutiques conduisant chaque année à la mort de millions de patients dans le monde (Wang et al., 2019).

De manière générale, l'analyse d'une pathologie spécifique nécessite souvent l'identification d'un grand nombre de patients. Jusqu'à présent, l'identification de ces patients était une tâche importante et chronophage, nécessitant une interprétation manuelle des dossiers médicaux et nécessitant des ressources considérables. Il s'agit d'une étape critique puisque 80% des informations cliniques pertinentes sont contenues dans le texte

3. ConSoRe : **Con**tinuum **So**in **Re**cherche.
(http://www.unicancer.fr/recherche/consore-moteur-recherche-pour-big-data-en-cancerologie).
4. CIFRE : **C**onvention **I**ndustrielle de **F**ormation par la **RE**cherche.
(http://www.anrt.asso.fr/fr/le-dispositif-cifre-7844).

des dossiers de santé [3].

Cette thèse s'inscrit dans le cadre du projet *ConSoRe*, et vise à créer des cohortes de patients résistants ou insensibles aux traitements anticancéreux grâce à l'utilisation et au croisement des connaissances et des informations contenues dans les comptes rendus médicaux. Pour ce faire, nous nous appuyons sur la mise en œuvre d'approches bio-informatiques faisant appel à des méthodologies basées, entre-autres, sur l'intelligence artificielle, l'apprentissage statistique et l'enrichissement sémantique.

La finalité de notre recherche est d'améliorer la modélisation des problématiques de résistances aux traitements d'oncologie par une contextualisation et l'enrichissement des données cliniques, en particulier les comptes rendus issus des dossiers patient. Pour cela, nous allons dans un premier temps exploiter l'ensemble des données disponibles afin de développer des outils capables de reconstruire le parcours néoplasique des patients. Cela va aider les médecins à améliorer les soins de santé et les accompagner au quotidien dans leur prise de décisions médicales. Dans un second temps, nous allons nous concentrer sur l'identification et la modélisation des résistances chez les patients cancéreux. Cette modélisation va nous permettre d'accroître nos connaissances sur le cancer de manière générale, et d'approfondir notre compréhension des phénomènes de la résistance aux traitements en particulier. L'identification des résistances aux traitements facilite la construction des modèles de prédiction afin de détecter des éventuels risques qui pourraient survenir pendant le traitement des patients, et prépare par ailleurs l'individualisation et le renforcement de la prévention en fonction du niveau de risque estimé. Cette démarche s'inscrit dans le cadre d'une médecine de précision, permettant de proposer de nouvelles solutions thérapeutiques adaptées à la fois aux caractéristiques de la maladie (cancer) et aux profils des patients identifiés. L'exploitation des cohortes de patients résistants aux traitements au travers de différentes études rétrospectives en suivant parallèlement l'évolution temporelle des pathologies de ces patients et les traitements qui leur ont été administrés, permet de guider la médecine vers des axes thérapeutiques novateurs. Faire valoir les connaissances contenues dans les comptes rendus de manière souveraine rejoint parfaitement les concepts de la médecine 4P qui se veut : ***P**rédictive, **P**réventive, **P**ersonnalisée et **P**articipative.*

Pour les besoins de nos recherches, trois types de cancer sont ciblés : le cancer du sein, le cancer du pancréas et le cancer du poumon. Ces trois cancers sont représentatifs des mécanismes de résistance aux traitements en oncologie, et offrent la diversité et la richesse nécessaires pour permettre une extension ultérieure à d'autres indications cancéreuses.

1.3 Problématique

1.3.1 Hétérogénéité des données

Le projet *ConSoRe* réunit plusieurs établissements de santé de la fédération UNICANCER [5], qui regroupe les 20 Centres de Lutte contre le Cancer en France (CLCC).

5. Fédération nationale des centres de lutte contre le cancer. (http://www.unicancer.fr/).

Les connaissances scientifiques et cliniques, ainsi que les données anonymisées des patients sont mises à notre disposition par 9 de ces centres cliniques. Ces données constituent la principale forme de communication entre les professionnels de la santé, et sont principalement enregistrées en texte libre pour faciliter leur utilisation par les médecins, plutôt que pour des considérations d'analyse des données. L'ensemble des documents est réparti sur deux catégories : des documents textuels structurés tels que les actes d'état civil, les PMSI (Programme de Médicalisation des Systèmes d'Information), les fiches de tumeur ou de chimiothérapie, etc., et des documents textuels non structurés. Ces derniers représentent la grande partie des documents, et sont représentés sous forme de comptes rendus de différents types : rapports de consultation, rapports d'imagerie, résultats d'analyses, etc.

Dans cette thèse, nos travaux s'articulent principalement autour du traitement des comptes rendus médicaux. Ces documents contiennent diverses informations médicales telles que : les traitements et les médicaments administrés aux patients, les actes d'interventions chirurgicaux, les antécédents personnels et familiaux (qui comprennent, entre-autres, les chirurgies antérieures, les hospitalisations et les maladies chroniques chez les membres de la famille), et les conclusions de fin de consultation, etc. Par ailleurs, ces documents contiennent également des informations administratives telles que celles relatives à l'établissement, par exemple : le nom de l'établissement, les adresses, les numéros de téléphone, etc.

1.3.2 Qualité des données

En dépit des efforts de quelques centres consacrés à la saisie des données cliniques dans un format structuré, la plupart des comptes rendus sont désormais conservés en texte libre pour permettre aux médecins de transcrire facilement des informations souvent pertinentes pour la recherche et les résultats cliniques. Ces comptes rendus contiennent un jargon spécifique, et ne suivent généralement pas de règles grammaticales formelles. Ils sont souvent rédigés dans un style orienté mots-clés et mis en forme avec de nombreux sauts de ligne, espaces blancs, listes à puces ou de comptage, etc. De plus, en raison des différences intrinsèques de styles de mise en forme qui varient selon les médecins et les établissements de soins, ces données conduisent à des problèmes naturellement hétérogènes principalement représentés par des incohérences dans les attributs des données. Ainsi, les tâches de pré-traitement de textes élaborées en amont de la fouille de texte telles que la tokenisation, la segmentation des textes en phrases et la normalisation de la variation des différents termes sont de plus en plus laborieuses.

Par ailleurs, le volume important et croissant de ces comptes rendus entraîne une augmentation des sources potentielles d'erreurs. En effet, ces documents contiennent des fautes d'orthographe (inversion des caractères, suppression des apostrophes, langage des abréviations, etc.) essentiellement à l'origine des non-mots absents du lexique. De plus, chaque centre possède son propre système d'information et conserve ses données dans un ou plusieurs formats : *PDF*, *MS Word*, *HTML* ou *Texte brut*. Le traitement de ces différents formats peut engendrer des défis supplémentaires pour le traitement des données "non propres". Par exemple, l'extraction des phrases issues d'un fichier *PDF* océrisé génère parfois des phrases mal découpées ou représentées par une suite

insignifiante de caractères spéciaux. Cela réduit considérablement les performances des tâches de traitements des données textuelles. Il est donc nécessaire de nettoyer ces données afin de garantir une recherche et une extraction d'informations pertinentes.

1.3.3 Les axes de recherches

Dans le cadre de notre recherche, le traitement des comptes rendus médicaux et l'exploitation des informations qui y sont contenues doivent permettre une compréhension approfondie des liens qui peuvent exister entre la maladie (cancer), son évolution et les traitements administrés aux patients. Les différentes tâches de fouille de texte permettent de valoriser et de structurer les connaissances contenues dans ces comptes rendus. Ces connaissances sont ensuite enrichies à l'aide des concepts extraits à partir des données structurées et des référentiels utilisés. Ce processus constitue une étape fondamentale dans l'acquisition de données "propres" indispensable à l'optimisation de l'identification des situations cliniques recherchées, par exemple, la création de cohorte de patients homogènes.

Pour mieux analyser ces documents, l'utilisation des techniques de traitement automatique du langage naturel (TALN) devient une nécessité. Ces techniques permettent d'automatiser plusieurs tâches grâce à la construction de représentations formelles qui nous permettent d'apporter des réponses précises à des besoins spécifiques. Ces techniques ont démontré à plusieurs reprises leur pertinence pour débloquer des preuves enfouies dans les rapports cliniques (Trivedi et al., 2017). Il ne s'agit donc pas simplement de sélectionner un fragment brut du texte, mais de mettre des éléments en relation pour restituer une information complète et structurée. Ces techniques sont réparties généralement sur plusieurs tâches subalternes allant de la préparation des données textuelles comprenant la tokenisation, la détection des limites des phrases, lemmatisation, etc., à l'extraction et la recherche d'informations à l'aide de modèles de reconnaissance d'entités nommées ou de classification automatique entre-autres en passant par les méthodes d'enrichissement sémantique telles que la désambiguïsation sémantique.

Par ailleurs, les techniques d'apprentissage automatique facilitent le développement d'outils de traitement automatique du langage naturel, notamment dans le domaine médical. En effet, ces dernières tirent leur puissance de leur capacité à fournir une analyse et une interprétation pertinentes quand il s'agit de traiter de grandes quantités de données. Les progrès réalisés dans le traitement des données médicales sont en grande partie associés à l'utilisation des algorithmes d'apprentissage automatique dans les tâches de traitement automatique de langage naturel. Cela permet par exemple que de grandes quantités de données patient puissent être facilement structurées et classées.

1.4 Contributions

Eu égard à ce qui précède, la réponse à la problématique posée dans le cadre de cette thèse est assujettie à plusieurs besoins : (i) de nouveaux processus de pré-traitements adaptés à l'ensemble des données textuelles mises à disposition afin de faciliter l'accès aux informations pertinentes, (ii) des modèles de classification automatiques des textes

afin d'organiser, de trier et surtout de comprendre le sens des informations extraites, et (iii) une modélisation des résistances aux traitements à travers la structuration des concepts médicaux, notamment l'identification des réponses aux traitements.

Dans ce cadre, les contributions de cette thèse s'articulent autour de trois approches complémentaires pour l'identification et la caractérisation des patients résistants aux traitements d'oncologie.

Notre première contribution consiste en une suite de méthodes de pré-traitements de texte appliquées à des comptes rendus médicaux. Ces pré-traitements dépendent fondamentalement de la qualité et de la nature de textes traités. La première opération vise à collecter, nettoyer et transformer les données afin d'unifier les traitements effectués en aval. Ce nettoyage doit permettre d'éliminer les bruits et les parties les moins utiles du texte qui peuvent réduire les performances de tâches ultérieures. La deuxième opération consiste en une méthode de tokenisation en s'appuyant sur des référentiels médicaux afin d'intégrer l'ensemble des exceptions liées aux domaine médical. La troisième opération est la détection des limites de phrases, étape importante pour la segmentation des documents en phrases sémantiquement reliées. La quatrième opération consiste à normaliser les textes à l'aide de la lemmatisation, la racinisation et un processus de correction orthographique. La cinquième opération est la détection des expressions de temporalité dans le texte. Cette étape est cruciale notamment pour la reconstitution du parcours néoplasique des patients. La sixième opération est la désambiguïsation sémantique des concepts médicaux à travers l'enrichissement sémantique et la construction de modèles de reconnaissances d'entités nommées. Enfin, une dernière opération est effectuée pour la détection des négations et des incertitudes exprimées dans le texte. Cette étape joue un rôle clé notamment dans le traitement des données cliniques. En effet, les médecins utilisent souvent la négation pour exclure un diagnostic ou un traitement, et des formulations hypothétiques pour souligner la prudence avec laquelle ils souhaitent s'exprimer sur l'existence ou pas d'un événement.

Notre deuxième contribution est un système constitué d'une extraction de concepts médicaux, et d'une classification automatique des phrases extraites des comptes rendus. L'extraction des concepts médicaux est fondée sur des modèles de reconnaissance d'entités nommées. En revanche, la classification des phrases quant à elle est décomposée en deux étapes. D'abord, une phase de représentation et d'extraction des caractéristiques à partir des textes en s'appuyant sur le calcul des plongements de mots. Ensuite, un modèle de classification automatique de phrases basé sur une combinaison d'un réseau de neurones convolutionnel (CNN) et d'un réseau de neurones récurrent (RNN).

La troisième et dernière contribution est une approche originale de détection des résistances aux traitements anticancéreux réalisée en deux temps. Tout d'abord, une structuration au niveau document des concepts identifiés dans les travaux précédents, suivie d'une structuration au niveau patient pour la définition de concepts composites tels que les évolutions tumorales. Ensuite, une identification des réponses aux traitements et des toxicités, éléments indispensables à la modélisation des résistances aux traitements.

1.5 Organisation du mémoire

Ce mémoire est structuré en cinq grandes parties.

1. La première partie définit un état de l'art global des travaux réalisés, et est composé de trois volets. Tout d'abord, nous allons voir les approches utilisées dans le pré-traitement et la préparation des données textuelles. Ensuite, nous mettons l'accent sur les différents algorithmes de classification automatique de textes. Enfin, nous abordons les travaux liés à la détection des phénomène de résistance aux traitement en oncologie.
2. La deuxième partie est consacrée au contexte et aux détails techniques et conceptuels du projet *ConSoRe* dans lequel s'inscrit cette thèse.
3. Dans la troisième partie, nous présentons nos solutions aux différents problèmes liées aux pré-traitement des données textuelles. Pour cela, nous allons aborder les détails de chaque approche proposée et les résultats obtenus.
4. La quatrième partie concerne les travaux liés à la classification de textes et à l'extraction de concepts médicaux. Pour cela, nous allons présenter dans un premier temps notre modèle de calcul de vecteurs de mots, ensuite nous menons une étude comparative des différents algorithmes d'apprentissage classiques tels que les arbres de décision (DT), les machines à vecteurs de support (SVM) ou encore les k plus proches voisins (kNN), et des algorithmes d'apprentissage profond, notamment les réseaux de neurones convolutionnels (CNN) et les réseaux de neurones récurrents (RNN).
5. La dernière partie est consacrée à notre approche de modélisation des résistances aux traitements à travers l'identification et la structuration des réponses aux traitements et des toxicités.

Enfin, nous concluons ce mémoire et présentons un bilan général de l'ensemble de nos contributions en faisant apparaître les perspectives de recherche que nous envisageons d'étudier à l'avenir.

CHAPITRE 2

ÉTAT DE L'ART

Sommaire

2.1 Introduction

Ce chapitre est structuré en trois parties qui relèvent de sous-domaines distincts mais complémentaires. Il aborde les travaux pertinents en lien avec notre recherche afin de présenter le contexte et la terminologie nécessaires pour la bonne compréhension de cette thèse. Comme mentionné dans le chapitre précédent, le but de nos travaux est d'identifier et de caractériser les patients résistants aux traitements anticancéreux et cela en s'appuyant principalement sur les données des comptes rendus écrits par les médecins dans les différents centres de lutte contre le cancer. Extraire et faire valoir ces informations permet d'orienter les médecins sur plusieurs axes, notamment lors de la prise de décision médicale. En oncologie, il existe des critères d'éligibilité assez stricts pour sélectionner des patients aux essais cliniques ou à des études rétrospectives. Il faut donc disposer d'un système robuste d'extraction et de recherche d'informations exploitant les comptes rendus médicaux des patients.

Dans ce chapitre, nous présentons les notions liées aux techniques de fouille de textes et de traitement automatique du langage naturel, ainsi que les définitions fondamentales du domaine de l'oncologie, en particulier les réponses et les résistances aux traitements anticancéreux administrés aux patients.

2.1.1 Techniques de pré-traitement des données textuelles

Dans la première partie de l'état de l'art, nous allons voir dans la littérature les techniques d'analyse et de pré-traitement des textes qui rendent celles-ci exploitables par les algorithmes d'apprentissage automatique (*machine learning*). Le traitement des données textuelles, y compris les tâches de fouille de textes, est souvent composé d'une séquence d'étapes. La première étape est celle du pré-traitement des données. Celle-ci est fondamentale et constitue une étape cruciale puisque la performance et l'efficacité des algorithmes d'apprentissage automatique des textes reposent essentiellement sur la pertinence de ses différentes tâches subalternes. En effet, les techniques de pré-traitements sont utilisées pour transformer les données textuelles non structurées en un format intermédiaire, stocké dans des bases ou des entrepôts de données, et compatible avec les modèles d'apprentissage automatique. Cette transformation repose sur différentes techniques d'identification des données textuelles notamment des techniques de traitement automatique de langage naturel comme la tokenisation qui permet de transformer un document texte en un vecteur de mots ; la racinisation et la lemmatisation qui visent à associer différentes formes d'un mot à une seule représentation ; la suppression des mots vides consiste à éliminer les mots courants qui n'ont aucun pouvoir discriminant ; la détection des limites de phrases est une étape souvent très utile notamment pour les documents n'ayant peu ou pas de structure définie ; la correction orthographique nécessaire pour rectifier les erreurs de saisie par exemple et éviter la perte d'informations ; la désambiguïsation lexicale pour déterminer le sens de chaque mot en prenant en compte son contexte lexical ; l'enrichissement sémantique, etc. Le pré-traitement dépend beaucoup de la qualité des données collectées et du processus de récupération (manuelle ou automatique), mais aussi du contenu et du domaine d'application. Pour cela, il est important d'effectuer en amont une petite exploration des données pour mieux comprendre leur nature. À première vue, ces

techniques peuvent sembler triviales. Cependant, elles comportent de réels défis tels que la gestion des acronymes dans la tokenisation et la désambiguïsation ou encore la gestion des limites de phrases qui sert à segmenter les textes en différentes parties sémantiquement cohérentes. Toutes ces techniques seront discutées dans ce chapitre.

2.1.2 Classification automatique de texte

Dans la deuxième partie de l'état de l'art, nous allons étudier de manière approfondie les approches et les algorithmes de classification automatique de textes non structurés avant de proposer et d'entraîner nos modèles de classification. L'alternative à l'écriture manuelle de programmes est l'apprentissage automatique qui est un sous-domaine de l'intelligence artificielle (IA) et qui a connu un fort développement à partir des années 1990. Dans ce domaine, la classification des textes est une tâche d'apprentissage automatique supervisé dès lors que les données d'apprentissage sont au préalable annotées avant d'être transmises en entrée aux algorithmes. L'annotation des données textuelles est une tâche spécifiquement linguistique et consiste à attribuer manuellement à chaque phrase (ou texte) une catégorie prédéfinie généralement en fonction de son contenu sémantique. L'algorithme essaie alors d'approcher une fonction $g : X \to Y$ qui mappe chaque instance (texte) de document de l'espace d'entrée $x \in X$ à l'une des catégories prédéfinies $y \in Y$ sur la base d'un ensemble d'apprentissage de tuples (x, y).

Au cours de ces dernières années, il y a eu une croissance exponentielle du nombre de documents et de textes complexes qui nécessitent une compréhension plus approfondie des méthodes d'apprentissage automatique pour pouvoir les classer ou les catégoriser avec précision dans de nombreuses applications notamment celle du domaine médical. Ces méthodes ont obtenu des résultats satisfaisants dans le traitement automatique du langage naturel. Le succès de ces méthodes repose sur leur capacité à comprendre les relations non linéaires existantes au sein de ces données. Cependant, il existe dans la littérature un large éventail de structures et d'architectures ainsi qu'une multitude d'algorithmes très variés. Dans le cadre de cette thèse, nous avons étudié les méthodes d'apprentissage automatique qui sont le plus souvent utilisées telles que : les arbres de décision (DT) (Morgan and Sonquist, 1963), les k plus proches voisins (kNN) (Patrick and II, 1969), la classification Naïve Bayésienne (NB) (Edwards, 1986), les machines à vecteurs de support (SVM) (Vapnik, 1964), et les méthodes d'apprentissage profond telles que les réseaux de neurones récurrents (RNN) (Rumelhart et al., 1985) et les réseaux de neurones convolutionnels (CNN) (LeCun et al., 1989). Les performances de ces algorithmes sont étroitement liées à la nature et le contexte des données traitées. La plupart des approches de classification de textes peuvent être décomposées en quatre phases : l'extraction de caractéristiques(Lewis, 1992), la réduction de dimensions (Stone, 1986), la sélection du classificateur (Giacinto and Roli, 1999) et l'évaluation (Gordon and Desjardins, 1995). Dans ce chapitre, nous allons discuter les différentes techniques de classification de texte abordées dans cette thèse.

2.1.3 Modélisation de la résistance et des réponses aux traitements en oncologie

La dernière partie de l'état de l'art est consacrée aux définitions des concepts médicaux pour la modélisation des résistances chez les patients atteints du cancer, et aux

travaux qui ont abordé les notions de *réponse*, *non réponse*, *sensibilité au traitement*, et les *toxicités*. Malgré les progrès de la détection et de l'amélioration de prise en charge du cancer, et l'augmentation significative de l'arsenal thérapeutique, les phénomènes d'échappement et de résistance aux traitements restent des problématiques majeures conduisant chaque année aux décès de nombreux patients. Dans cette thèse, l'identification de cohortes des patients résistants au traitement d'oncologie permet d'accompagner les praticiens fondamentalement sur trois axes. D'abord, construire des modèles prédictifs pour la détection précoce des patients susceptibles de développer des résistances ou des insensibilités au traitement, et cela relève de la prévention. Ensuite, permettre au médecin de mener des suivis longitudinaux et des études rétrospectives sur ces cohortes de patients afin de suivre l'évolution de la maladie (naturelle ou sous traitement) et de comprendre et d'identifier les facteurs et les éléments déclencheurs de ces phénomènes de résistance. Cela passe par le développement d'outils informatiques à visée diagnostique et de recherche. Enfin, permettre aux laboratoires et aux industries pharmaceutiques de développer des molécules adaptées à la fois aux caractéristiques de la maladie et aux profils des patients concernés. Cette démarche s'inscrit dans le cadre d'une médecine de précision (ou personnalisée). La nécessité de stratification des populations de patients pour apporter une solution thérapeutique adaptée au plus près des besoins médicaux se traduit par l'utilisation des outils de théranostique. Ils sont devenus essentiels à la stratégie de recherche et de développement des médicaments. Ce dernier chantier ne fait pas partie de notre périmètre d'étude.

2.2 Préparation et pré-traitement des données textuelles

2.2.1 Introduction

Avec le développement de l'informatique, nous constatons une augmentation sans précédent de la production et du stockage des données en général, et textuelles en particulier. En effet, Gantz and Reinsel (2011) affirment que l'information numérique produite dans le monde aurait plus que doublé entre 2010 et 2012, passant de $1,2$ à $2,8$ zettaoctets [1]. En fait, cette tendance va en s'accentuant, elle passera à $175Zo$ en 2025. Bien que l'aspect volumétrique des données textuelles semble être maîtrisé, sa dimension hétérogène reste un défi pour la communauté scientifique. Ces masses de données textuelles peuvent présenter une grande diversité et de nombreuses caractéristiques qui entravent les opérations d'analyse, et engendrent un problème difficile lié à leur traitement automatique. Par conséquent, il est difficile d'extraire des connaissances de manière rapide et efficace à partir de ces données dans leur état brut.

Le traitement des données textuelles nécessite des algorithmes intelligents permettant de sélectionner rapidement des informations pertinentes à partir des grandes quantités de données afin de répondre à une requête d'un utilisateur. Ce processus s'appelle fouille de texte (*text mining*), et est souvent appliqué à des données textuelles non structurées ou semi-structurées. La fouille de texte permet de récupérer le contenu d'un texte, mais aussi le contexte dans lequel il est cité.

1. $1Zo = 10^{21}$ octets

Cependant, en raison d'un problème de surabondance de modèles, il est difficile d'identifier les seuls ensembles de résultats pertinents. Pour rendre cela possible, les données doivent être transformées en un format pivot — une structure plus appropriée pour faciliter l'extraction des informations pertinentes — en s'appuyant sur diverses techniques de raffinement afin que les algorithmes d'apprentissage automatique puissent être appliqués d'une manière efficace. Cette transformation est appelée : prétraitement des données ou encore préparation de données, et est le préalable de tout travail sur un corpus de données textuelles. Le pré-traitement des textes est une étape essentielle dans tout système de traitement automatique de langage naturel, puisque les caractères, les mots et les phrases identifiés à ce stade sont les unités fondamentales à toutes les étapes ultérieures dans le traitement de données textuelles. Compte tenu des spécificités et de la richesses des formats que peuvent contenir les données textuelles, par exemple les formats numériques ou les formats de dates pour l'expression de la temporalité, ou encore les signes de ponctuation qui ne sont parfois pas susceptibles d'aider la fouille de texte, ces opérations de pré-traitement permettent d'effectuer une synchronisation massive des corpus prêts à être exploités par des applications de fouille de texte. Dans la littérature, la communauté de recherche a travaillé sur différentes techniques de pré-traitement qui sont toutes différentes. Certaines de ces techniques sont discutées ici.

2.2.2 Travaux connexes

2.2.2.1 Travaux connexes dans le domaine général

Dans la littérature, de nombreux chercheurs ont abordé la question des pré-traitements appliqués à des données textuelles afin de les rendre compatibles avec les algorithmes d'apprentissage automatique. Ainsi, plusieurs outils ont vu le jour, par exemple : *Unitex* (Paumier, 2003), *GATE* (Cunningham et al., 2002) et *LinguaStream* (Widlöcher and Bilhaut, 2006). Cependant, différentes expériences nous ont montré les limites de ces outils. En effet, certains ne disposent pas d'une cartouche de correction orthographique, tandis que d'autres sont très spécialisés sur certains types de données et donc très peu ou pas généralisables.

(Wang, 2004) décrit cinq approches différentes permettant, à partir de textes bruts, la création d'une structure intermédiaire pour la fouille de texte. La première approche consiste à sélectionner un ensemble de mots à partir d'un document pour le représenter. De nombreux travaux (Ahonen-Myka, 1999a, 2002, 1999b; Li and Chung, 2005) ont utilisé cette approche. Pour obtenir de meilleures performances, les mots vides peuvent être supprimés de cet ensemble de mots. La deuxième approche affine chaque sac de mots obtenu en référençant une liste de mots clés (Calvo-Flores et al., 2002; Feldman et al., 1998; Feldman and Hirsh, 1996). Ces derniers sont tirés avec différents schémas tels que le schéma de la pondération basé sur la fréquence (Calvo-Flores et al., 2002) ou l'extraction de mots clés basée sur les CRF (champs aléatoires conditionnels) (Zhang et al., 2008). En revanche, la suppression d'informations riches pourrait impacter la fouille de texte, ce qui représente un inconvénient dans cette approche. La troisième approche est celle du document prototypique (Rajman and Besançon, 1998) et qui est composée de deux modules. Le premier est le *Part-of-Speech Tagging* (POS)(Feldman

et al., 1998) qui attribue des tags de manière automatique à des mots en fonction du contexte grammatical de la phrase. Le second consiste à extraire les termes (extraction d'entités nommées) qui dépendent du domaine d'application (Lin et al., 1998). Ces termes peuvent être des phrases ou de simples mots. La quatrième approche : textes à multi-termes (Heinonen et al., 1999) consiste en la co-occurrence d'un ensemble de mots dans des données brutes. La dernière approche décrite est conceptuelle et extrait d'abord les termes clés et leur relation syntaxique afin de pouvoir extraire des concepts plus significatifs sur le plan sémantique.

D'autres travaux se sont focalisés sur des méthodes de pré-traitement de textes moins généralistes et plus spécialisées. Hotho et al. (2005) a identifié trois de ces étapes. La première est la *Tokenisation*, qui a pour but de générer un flux de mots en supprimant les suites de caractères vides et les signes de ponctuation. Le document résultant est appelé dictionnaire. La deuxième étape cherche à réduire la dimensionnalité du problème et consiste à appliquer des opérations telles que le Filtrage (suppression des mots vides) ; la *Lemmatisation* (mapper les mots nominaux à leur forme singulière et les verbes à leur forme à l'infinitif). Cette opération est sujette aux erreurs. Et enfin la *Racinisation* (pour ramener chaque mot à sa racine, (Porter, 1980)). La troisième étape est également utilisée pour réduire davantage les mots du lexique et consiste à sélectionner les mots clés. Hotho et al. (2005) a utilisé l'entropie à cette fin de manière à estimer l'importance d'un mot en calculant l'entropie. Une autre technique de sélection des mots clés est le calcul des distances entre les couples mots d'un corpus afin de mesurer leur similarité. La plupart des mots clés associés ayant une distance minimale sont sélectionnés (Kardan et al., 2013).

2.2.2.2 Travaux connexes dans le domaine médical

Dans le traitement des données textuelles du domaine médical, Mathiak and Eckstein (2004) ont défini une méthode de fouille de texte composée de cinq étapes. Le pré-traitement des données est l'une de ces étapes durant laquelle plusieurs opérations ont été définies : la tokenisation, Le *Part-Of-Speech Tagging*, le calcul des fréquences des termes et enfin les méthodes de racinisation. En utilisant ce processus, ils ont montré qu'il est possible de réduire considérablement le temps en extrayant les termes les plus intéressants dans une démarche de fouille de texte.

Dans l'article (Sun et al., 2018), les chercheurs ont montré l'importance et l'utilité des techniques de pré-traitement en amont de la fouille de texte et en particulier quand il s'agit d'explorer des données semi-structurées ou non structurées telles que le texte médical. Les auteurs attestent qu'il est nécessaire de pré-traiter les données sources afin d'améliorer la qualité des données, et ainsi améliorer les résultats des tâches ultérieures. Par ailleurs, ils montrent qu'il est primordial d'adapter le choix des techniques de traitement en fonction des types des données à traiter.

Dalianis (2014) trouve que les textes cliniques contiennent souvent du bruit ce qui les rend difficiles à traiter, et que les technologies standard de recherche d'information (RI) ne peuvent pas être utilisées pour récupérer des informations pertinentes à partir de ce type de textes. Ces derniers doivent être pré-traités à l'aide de : la vérification

ou correction orthographique grammaticale, l'extension des abréviations, la lemmatisation et enfin la détection de la négation, afin de normaliser ces textes.

2.2.3 L'utilité des pré-traitements

La fouille de données textuelles est une tâche souvent complexe qui doit se baser sur la modélisation informatique des données, mais aussi l'étude linguistique. En effet, les scientifiques sans formation linguistique, sont souvent surpris par l'éventail des possibilités linguistiques pour exprimer même les concepts les plus simples (Cohen and Hunter, 2008). Par ailleurs, il est plus judicieux de prendre en compte la langue utilisée et le domaine d'application entre autres lors de l'exécution de ces techniques notamment le pré-traitement des données. En fonction de l'exigence de l'application, certaines étapes de pré-traitement sont appliquées ou pas.

(Torunoğlu et al., 2011) ont étudié l'impact du pré-traitement sur deux grands ensembles de données provenant de journaux turcs. Ils ont utilisé des techniques de pré-traitement basiques comme la suppression des mots vides, la racinisation, la pondération des termes. Ils ont montré que la racinisation a un grand impact sur l'Extraction d'Information (*Information Retrieval*). En revanche, la suppression des mots vides et la racinisation ont moins d'impact sur l'exactitude de la classification. Un travail similaire a également été réalisé par (Sohail and Hassanain, 2012) sur des données en langue arabe. À cet égard, ils ont présenté une étude approfondie sur les différentes techniques de pré-traitement que les chercheurs utilisent pour gérer ce type d'applications de pré-traitement de texte basées sur la langue arabe.

Dans l'article de (Etaiwi and Al-Naymat, 2017), les chercheurs ont étudié l'impact d'utilisation des techniques de pré-traitement de textes sur la précision de détection en utilisant des algorithmes d'apprentissage automatique tels que SVM (Vapnik, 1964) et NB (Edwards, 1986). Dans leur analyse, ils ont utilisé un nombre limité de caractéristiques linguistiques qui réduisaient la précision de la détection. Ils ont conclu que la suppression des signes de ponctuation, les mots vides et les caractères numériques permet de mettre l'accent sur les mots importants, et donc d'optimiser les résultats de toute approche de traitement de texte.

En dépit des différences des approches citées dans les travaux ci-dessus, elles démontrent que la plupart des techniques de fouille de texte sont fondées sur l'idée qu'un document texte peut être présenté par un ensemble de mots. Si chaque document doit être considéré comme un vecteur linéaire, alors pour chaque mot du document, une valeur numérique est stockée montrant son importance dans le document (vecteur linéaire). Les recherches ci-dessus effectuées dans le domaine du pré-traitement indiquent que diverses méthodes découvrent progressivement un lot représentatif de mots sur lesquels des techniques de fouille de texte peuvent être utilisées pour découvrir des modèles intéressants de cognition.

2.2.4 Considérations conceptuelles et empiriques

À ce stade, nous avons pu voir l'importance du pré-traitement des textes dans la fouille de texte. Cependant, le pré-traitement peut également supprimer des informations utiles (lors de la suppression des mots vides), introduire des erreurs dans l'analyse (lorsque la racinisation confond des mots sémantiquement distincts) et donc modifier radicalement les performances des tâches ultérieures (Boyd, 2016). En d'autres termes, le pré-traitement peut affecter la fiabilité et la validité des résultats de la même manière que les degrés de liberté des chercheurs (Simmons et al., 2011) peuvent modifier les résultats de la recherche pendant le codage méta-analytique (Wanous et al., 1989), la recherche qualitative (Jonsen et al., 2018) et le dépistage des réponses quantitatives (Meade and Craig, 2012). De plus, même si la plupart des recherches sont d'accord sur l'importance du pré-traitement des textes dans les applications de fouilles de texte, nous avons pu constater que les chercheurs fournissent des recommandations de pré-traitement parfois contradictoires.

Nous allons aborder dans ce qui suit les techniques de pré-traitement de textes de manière plus approfondie afin d'observer ces recommandations contradictoires et fournir des lignes directrices pour la recherche en fouille de texte. Mais dans premier temps, il est nécessaire de comprendre les considérations conceptuelles et empiriques qui devraient guider les choix de certaines techniques de pré-traitement ; c'est-à-dire, pourquoi et quand le pré-traitement peut affecter les résultats de la fouille de texte étant donné que ce pré-traitement doit être considéré dans le contexte de son corpus.

Afin de mieux répondre à une requête dans un système de fouille de texte, il faut comprendre trois éléments. Premièrement, ce qui est véhiculé comme information et de quelle façon : le contenu et le style des textes (Pennebaker et al., 2003). En effet, le pré-traitement permet de déterminer si des éléments de style qui peuvent être informatifs pour des questions précises, tels que l'utilisation d'abréviations ou les fautes d'orthographe, sont présents dans le texte. Deuxièmement, les caractéristiques des textes : la taille du corpus et la longueur moyenne d'un document car les grands corpus ont le pouvoir de fines distinctions entre, par exemple, les versions singulier et pluriel des noms. Ainsi, le choix de la lemmatisation et la racinisation est étroitement lié à la taille du corpus (Kern et al., 2016 ; Kobayashi et al., 2017b). Troisièmement, comme le note le manuel opérateur de l'outil *Linguistic Inquiry and Word Count* (LIWC), lors du pré-traitement, il faut "garder à l'esprit quels sont vos objectifs en analysant les données" (Pennebaker et al., 2015). En d'autres termes, il faut tenir compte des étapes ultérieures du processus de fouille de texte, notamment, la nature des questions auxquelles on veut répondre.

2.2.5 Pré-traitements

Nous allons à présent expliquer les techniques de pré-traitement et leurs effets sur la fouille de texte, en les reliant aux considérations conceptuelles et empiriques susmentionnées.

2.2.5.1 Collecte, nettoyage et transformation des données

La première étape dans toutes les applications de traitement automatique de langage naturel est la collecte des données. Ces dernières peuvent provenir de sources diverses, et elles sont soit structurées soit non structurées. Il est souvent question de concevoir un lac de données (*Data Lake*) afin de maintenir les sources de données dans leur forme originelle pour éviter ou minimiser la perte d'informations au moment de leur traitement. Une autre solution consiste à concevoir un multi-stores permettant de regrouper les sources de données hétérogènes, stockées sous des formats différents : *PDF*, *MS Word*, *Texte brut*, *HTML*, etc. Afin d'extraire les textes bruts, différentes techniques sont utilisées telles que : parser les documents (*PDF* et *MS Word*) afin d'extraire le texte brut ; retirer les balises *HTML*, etc.

Les textes bruts contiennent d'imperfections et parfois des caractères issus des erreurs de retranscription. Les nettoyer doit permettre d'éliminer les bruits et les parties non significatives qui peuvent réduire les performances des tâches de fouille de texte, tout en conservant les caractéristiques textuelles essentielles.

2.2.5.2 Tokenisation

La tokenisation est le processus qui consiste à diviser un document texte en unités de traitements significatives (mots) appelés jetons (*tokens* en anglais) (Hassler and Fliedl, 2006). Ce processus est fondé sur un modèle spécifique en utilisant des règles et des expressions régulières, et utilise les signes de ponctuation et l'espace blanc avec quelques heuristiques pour éviter les cas non triviaux. Il s'agit généralement du premier processus de toute application de traitement automatique de langage naturel. Cette opération a trois principaux objectifs : le premier est d'obtenir tous les mots du corpus ; le deuxième est d'identifier des mots-clés significatifs ; et le troisième est de reconnaître les limites des phrases et des mots. La tokenisation est généralement considérée comme une tâche triviale, en particulier lorsqu'on considère des langues telles que le français où les mots sont séparés par des espaces blancs. Cependant, il existe de nombreux obstacles à gérer tels que les acronymes, les abréviations ou encore les exceptions reliées au domaine d'application (par exemple, les médicaments ou les noms de traitements dans le domaine médical). En effet, le point « . » indique généralement la fin d'une phrase, et est considéré comme un jeton à part entière. Cependant, lorsqu'il apparaît dans des abréviations ou des acronymes, il devient une partie intégrante du jeton. La tokenisation joue un rôle important dans l'analyse lexicale, mais aussi dans la segmentation des textes en phrases, comme nous allons le voir dans ce chapitre. Une autre difficulté liée à ce processus est la gestion des césures pour les mots contenant des traits d'union.

De nombreuses méthodes de tokenisation (Attia, 2007; Huang et al., 2007; Labadié and Prince, 2008) sont proposées en fonction des différentes langues. Chaque langue a des spécificités qui doivent être considérées indépendamment. En effet, pour le chinois, le japonais et le thaï, le texte est représenté comme une séquence de caractères sans espace entre eux. Pour obtenir des mots, des méthodes particulières doivent être appliquées. Pour l'arabe, les pronoms sont souvent attachés aux verbes. La tokenisation est une étape importante surtout quand il s'agit de traiter des documents issus

d'un système OCR (Reconnaissance Optique de Caractères), qui pourraient se montrer comme une source de bruit, pour identifier les mots (Kolak et al., 2004). Dans cette thèse, compte tenu des différentes spécificités liées au choix de la langue utilisée, nous nous intéressons en particulier aux travaux de tokenisation réalisés sur des documents de texte rédigés en langue française.

2.2.5.3 Normalisation

La normalisation de texte fait référence à un ensemble de tâches qui transforment le texte en une forme canonique plus standard. C'est l'une des étapes clés du traitement automatique de textes dans lesquels la variation orthographique abonde de la norme contemporaine. La normalisation des textes a de nombreux objectifs. Elle permet de réduire les variations des variables à traiter. Cela est particulièrement utile dans les systèmes de questions/réponses qui peuvent être conçus de manière à interroger une seule variante standard mais qui prend en compte toutes ses variantes orthographiques (dialectales, familières ou argotiques). La normalisation permet de réduire la dimensionnalité du problème lorsqu'il s'agit d'utiliser des algorithmes d'apprentissage automatique. En d'autres termes, il s'agit de réduire le nombre de vecteurs représentant les mots du corpus dans un espace vectoriel.

En fonction de l'application de fouille de texte et du niveau de sa représentation, différentes techniques peuvent être appliquées. En plus de la racinisation et de la lemmatisation, certains travaux recommandent de convertir tout le texte en une seule casse en minuscules ou en majuscules (Banks et al., 2018; Kobayashi et al., 2018). En revanche, cela pourrait compliquer davantage les tâches de traitement de textes telles que la désambiguïsation sémantique ou la détection des limites de phrases. La conversion en minuscules n'est pas toujours nécessaire puisque, comme on l'a précisé, il faut prendre en compte les exigences conceptuelles et les objectifs à tenir. Par ailleurs, d'autres recherches recommandent de supprimer les caractères non alphabétiques (Banks et al., 2018) du texte. Néanmoins, la ponctuation est un élément du style qui peut se montrer nécessaire dans les tâches ultérieures.

2.2.5.4 Lemmatisation et Racinisation

La lemmatisation est une technique qui effectue une analyse morphologique complète d'un mot et produit également la forme de base du mot en s'appuyant généralement sur des dictionnaires. En d'autres termes, les mots sont confrontés à un dictionnaire qui sert à faire correspondre les variantes d'un mot à sa forme de base. La lemmatisation dépend souvent du langage utilisé (Habash et al., 2009; Suhartono, 2014). De façon générale, un lemmatiseur comporte trois parties : (i) un ensemble de règles, (ii) un lexique de mots de base, et (iii) un algorithme de lemmatisation. Dans la littérature, certains travaux ont abordé la lemmatisation avec des approches fondées sur les règles (Plisson et al., 2004) tandis que d'autres ont tenté des approches d'apprentissage automatique. Dans (Jongejan and Dalianis, 2009), les auteurs ont proposé une approche supervisée pour apprendre automatiquement les règles de lemmatisation des langues néerlandaise et allemande. À l'origine, le but principal de la lemmatisation est de réduire le temps de calcul et d'améliorer le rappel dans la recherche d'informations.

Cependant, ces techniques peuvent également augmenter la puissance en réduisant la dimensionnalité.

La racinisation est le processus de réduction des formes infléchies d'un mot à leur racine. Dans le traitement automatique de langage naturel, de nombreux algorithmes de dérivation ont été proposés (Jivani et al., 2011; Paik et al., 2011; Majumder et al., 2007). Ils consistent à supprimer la fin des mots pour ne garder que la racine. De cette manière, l'utilisation de la racine est appliquée pour réduire la taille de l'ensemble des données. Dans la littérature, il existe trois approches de base pour la racinisation. La première consiste à supprimer les préfixes et suffixes des mots pour ne garder que la racine (Dawson, 1974; Lovins, 1968; Paice, 1990; Porter, 1980). La deuxième approche est statistique et ne dépend pas de la langue (Rogati et al., 2003; Majumder et al., 2007). La troisième approche est mixte, et comprend des algorithmes de flexion et d'autres basés sur un corpus.

La prudence est de mise car l'utilisation de ces deux techniques peut augmenter la couverture des dictionnaires, mais si cela introduit des erreurs ou confond des mots distincts, cela réduira la précision et la validité de la fouille de textes. En effet, la lemmatisation d'un grand dictionnaire est une stratégie très chronophage. De plus, elle peut être une source d'ambiguïté quand il s'agit de traiter des polysèmes (mots ayant plusieurs sens). Par ailleurs, le regroupement des variantes de mots en une seule racine peut supprimer les différences de style du texte. La racinisation utilise souvent des heuristiques basées sur des règles pour supprimer les suffixes des mots (Jivani et al., 2011), ne laissant que la racine, sans tenir compte des homographes (les mots écrits de la même manière mais avec des significations différentes). Cette technique peut aussi associer des mots différents à la même racine. De plus, cela peut créer des mots irréels ce qui les rend difficile à interpréter.

2.2.5.5 Correction orthographique

La correction orthographique permet de détecter les mots mal orthographiés et de proposer si possible une correction. Cette tâche de normalisation est souvent très utile dans le traitement de données textuelles, surtout quand il s'agit de données retranscrites par un processus de reconnaissance optique de caractères (OCR). Les fautes de frappe n'apparaissent que quelques fois dans un corpus et ne fournissent probablement que peu de valeur pour une analyse ultérieure dans la fouille de texte. Leur correction a le potentiel de réduire la dimensionnalité du problème. Les quatre types d'erreurs les plus fréquemment rencontrés sont : l'insertion, la suppression, la substitution et la transposition. Ces types englobent environ 80% de toutes les fautes d'orthographe (Kukich, 1992). La plupart des vérificateurs orthographiques utilisent des techniques de correspondance de chaînes avec un dictionnaire spécifique ou un thésaurus pour une langue donnée. La distance d'édition, aussi appelée distance de *Damerau-Levenshtein* (Kukich, 1992), permet de calculer le nombre minimum d'opérations nécessaires pour transformer une chaîne de caractères en une autre chaîne, et est souvent utilisée dans les systèmes de correction orthographique. D'autres types d'erreurs existent telles que les erreurs non lexicales, c'est-à-dire des erreurs qui aboutissent à un mot qui existe dans le dictionnaire et qui ne peuvent donc pas être détectées et corrigées sans prise

en compte du contexte environnant. D'un autre côté, les fautes d'orthographe peuvent s'avérer informatives pour des questions précises et représenter un style de texte en particulier. La nécessité d'effectuer ce travail de pré-traitement dépend en réalité des besoins spécifiques liés à la tâche de fouille de texte.

2.2.5.6 Suppression des mots vides

Les mots vides sont des mots utiles pour faire des phrases, mais qui ont peu de contenu informatif. Ce sont des mots très courants dans un corpus et peuvent devenir non informatifs pour la compréhension du sens d'un document ou d'un texte. Ces mots sont souvent supprimés lors du pré-traitement pour la fouille de texte. Cela permet de réduire le temps de calcul, et d'améliorer les performances des tâches de classification de texte (Saif et al., 2014; Wilbur and Sirotkin, 1992; Yang and Wilbur, 1996; Silva, 2003; Makrehchi and Kamel, 2008). En effet, des mots comme « le » ou « et » sont des mots courants et n'ont pas de capacité de discrimination. Les chercheurs ont recommandé de toujours supprimer les mots vides (Banks et al., 2018), sauf pour les documents courts (Kobayashi et al., 2018).

La suppression des mots vides a été abordée de plusieurs manières. La première approche, et la plus utilisée, concerne la suppression en rejetant les mots apparaissant dans une liste compilée de mots vides dépendant de la langue. Certains schémas (Lo et al., 2005; Rose et al., 2010) sont également proposés pour la génération automatique de listes de mots vides. On peut également personnaliser une liste de mots vides en fonction des mots qui peuvent être intéressants pour une tâche souhaitée.

D'autres approches, dites automatiques, se contentent de supprimer les mots ayant atteint un certain nombre d'occurrences dans le document ou dans le corpus. Ce nombre d'occurrences est un seuil défini en fonction des exigences et du domaine d'application.

2.2.5.7 Détection des limites de phrases

La tâche de la détection des limites de phrases ou de la désambiguïsation des limites de phrases (*Sentence Boundary Detection*) est d'identifier les éléments de phrases dans un texte. La plupart des applications de TALN reposent sur cette tâche puisqu'elles prennent une phrase comme unité d'entrée. C'est le cas pour la traduction automatique (Wong et al., 2006), la recherche d'informations (Wang et al., 2012) et la reconnaissance d'entités nommées (NER). Il s'agit d'une tâche non triviale, car les erreurs engendrées par un mauvais découpage de phrases se propagent dans les tâches ultérieures du traitement de texte et auront un impact négatif sur leurs performances. La complexité de cette tâche est liée à l'ambiguïté potentielle des signes de ponctuation. En français, un point «.», qui est utilisé pour signaler la fin d'une phrase, peut également être utilisé pour désigner une abréviation, un acronyme, un point décimal, des séparateurs dans les adresses e-mail, etc. De plus, cette ambiguïté varie selon les différents styles de textes ou de corpus spécifiques, et selon la langue utilisée. Les performances élevées obtenues dans la segmentation de textes en phrases dans certains travaux font d'elle un problème résolu (Kiss and Strunk, 2006). Cependant, les bons résultats obtenus sur le domaine général ne se maintiennent pas toujours sur les données des domaines

spécialisés, notamment dans le domaine médical qui connaissent un style d'écriture et de formatage très particulier.

Dans la littérature, les approches de segmentations de texte en phrases se répartissent globalement en trois classes : (1) les approches basées sur les règles qui utilisent des heuristiques et des listes de caractères de ponctuation. En effet, de nombreux travaux ont tenté de résoudre cette question avec des approches fondées sur l'utilisation de ponctuations ambiguës pour déterminer si la ponctuation courante est un vrai délimiteur (Grefenstette and Tapanainen, 1994). (2) des approches basées sur l'apprentissage automatique entraînées dans une configuration supervisée et qui exploitent des algorithmes tels que les arbres de décision ou les réseaux de neurones (Palmer and Hearst, 1997). (3), des applications non supervisées de l'apprentissage automatique (Kiss and Strunk, 2006), qui ne nécessitent que des corpus bruts.

2.2.5.8 Désambiguïsation sémantique

La désambiguïsation sémantique est une technique qui permet de déterminer le sens des unités lexicales polysémiques dans un texte, et s'effectue en tenant compte du contexte. Elle consiste donc à sélectionner automatiquement le sens le plus approprié d'un mot en fonction du contexte dans lequel il est cité (Nancy and Jean, 1998; Navigli, 2009), et le distinguer des autres définitions qu'on peut attribuer à ce mot. La désambiguïsation sémantique n'est pas une fin en soi, mais constitue une étape « intermédiaire » indispensable pour la préparation des données textuelles aux différentes tâches ultérieures de traitement de texte (Kilgarriff, 1997; Navigli, 2009; Zhong and Ng, 2012). En effet, elle permet d'améliorer les performances et l'exactitude de différentes applications, en particulier celles qui visent à comprendre le texte en langage naturel, telles que la traduction automatique ou la recherche d'informations.

Du fait de cette grande variété d'intérêts, plusieurs chercheurs ont vite identifié les difficultés liées à la désambiguïsation sémantique. Toutefois, les approches proposées dans chaque domaine ont également été multiples et très diverses, en fonction des besoins et des savoirs afférents à chacune des matières concernées. Dans la littérature, parmi les méthodes connues de désambiguïsation sémantique (Navigli, 2009), il existe principalement trois catégories. La première regroupe les méthodes qui utilisent des règles simples pour détecter le sens d'un mot dans un ensemble préalablement défini (Hindle, 1989). La deuxième catégorie repose sur l'apprentissage automatique supervisé et l'utilisation des corpus de textes annotés réunissant des exemples d'instances désambiguïsées de mots (Bakx et al., 2006; Navigli, 2009). La dernière catégorie concerne les systèmes non supervisés qui utilisent des connaissances provenant des réseaux sémantiques (Schwab et al., 2012; Navigli, 2009).

2.2.5.9 Détection de la négation et de l'incertitude

La détection automatique de la négation et de l'incertitude fait partie des tâches de pré-traitement dans l'extraction d'informations, l'analyse des sentiments et le traitement automatique de langage naturel de manière générale. Une phrase est considérée comme négative, si elle contient une forme de négation sur une partie ou toute la

phrase. Une phrase est incertaine lorsqu'à sa lecture seule, on n'est pas en mesure d'affirmer la présence ou non des concepts qui y sont contenus. Ces définitions génériques n'indiquent aucun moyen pour permettre de l'identifier dans les textes, mais laissent supposer une grande variété de réalisations linguistiques.

Les phrases négatives et spéculatives ont un rôle important dans la langue car elles modifient souvent la polarité et donc le sens des phrases qui suivent dans un texte. Cependant, il est important de savoir les détecter car elles représentent l'une des sources d'erreur les plus répandues dans la fouille de texte (Schwartz et al., 2013). La gestion de négations et des incertitudes permet de capturer plus d'informations sémantiques, et par conséquent réduit les erreurs et augmente la validité des tâches ultérieures.

Une méthode rudimentaire de gestion des négations consiste à ajouter une chaîne de caractères prédéfinie (par exemple : «not_») à chaque mot qu'elle précède, créant ainsi un nouvel uni-gramme distinct (Speer, 2018). Cependant, la principale difficulté est d'être capable de résoudre les ambiguïtés, c'est-à-dire différencier les contextes où les termes sont utilisés comme marqueurs de ceux où ils ne le sont pas. De plus, ces marqueurs ne sont pas toujours présents sous formes de mots mais peuvent être des préfixes ou même des chiffres. L'effet d'un marqueur de négation ou d'une incertitude s'étend généralement sur toute la phrase, ou simplement sur une partie de cette dernière. Cette dernière est appelée portée (*scope*). Identifier automatiquement les portées, de manière précise, est une tâche complexe. En effet, la portée peut s'étendre des deux côtés d'un marqueur et peut être discontinue. De plus, plusieurs portées peuvent se chevaucher quand il y a plusieurs marqueurs dans une même phrase. Ces portées doivent être repérées indépendamment l'une de l'autre.

Dans la littérature, il existe généralement deux familles d'approches pour la détection automatique des phrases négatives ou spéculatives. Des travaux utilisant des règles à l'aide d'expressions régulières et de systèmes experts tels que *NegEx* (Chapman et al., 2001) et *Negfinder* (Mutalik et al., 2001) et des travaux récents qui se basent sur des méthodes de classification automatique supervisée (Lapponi et al., 2012; Read et al., 2012; Packard et al., 2014; Fancellu et al., 2016).

2.2.5.10 Étiquetage morphosyntaxique (POS)

L'étiquetage morphosyntaxique (aussi appelé étiquetage grammatical) consiste à *taguer* chaque mot dans un texte par sa catégorie lexicale correspondante, également appelée balise ou classe de mots, par exemple : verbe, adjectif, adverbe, préposition, etc. Cette opération doit être considérée comme un sous-processus de pré-traitement linguistique plutôt que comme un pré-traitement de texte (Hotho et al., 2005). L'étiquetage en domaine de spécialité est limité par la disponibilité d'outils et de corpus annotés spécifiques au domaine.

L'importance de l'étiquetage morphosyntaxique vient de la quantité d'informations qui pourraient être extraites sur un mot et ses voisins, et pourrait être exploité dans une logique de désambiguïsation sémantique ou de classification de texte en réduisant les erreurs d'interprétation de mots similaires.

Dans le marquage basé sur des règles, les balises sont attribuées à chaque mot en utilisant des règles généralement basées sur le cadre contextuel (Hasan et al., 2007). La précision de ces étiqueteurs n'est pas très bonne et ceux-ci ne sont pas robustes, tandis que les étiqueteurs basés sur des approches stochastiques ont de meilleures performances, et leur précision est plus élevée. La plupart des approches stochastiques sont basées sur le modèle de Markov (Hasan et al., 2007). Par ailleurs, il existe des approches hybrides qui bénéficient des avantages des deux méthodes évoquées.

2.2.5.11 La reconnaissance d'entités nommées (NER)

La reconnaissance d'entités nommées est une tâche souvent utilisée en amont de l'extraction d'informations (IE), et qui consiste à identifier et classer automatiquement les entités spécifiées dans un ou plusieurs corpus de textes dans des catégories préalablement définies, telles que : personnes, lieu, date, etc. Cette technique est utilisée dans de nombreux domaines de l'intelligence artificielle (Dalloux et al., 2021), notamment le traitement automatique de langage naturel et l'apprentissage automatique. Elle est utile dans diverses applications comme la classification automatique, la compréhension du langage naturel et les systèmes de questions/réponses.

Cette tâche a été abordée par deux approches principales. La première est une approche basée sur des règles (orientée connaissances), fondée sur un lexique et un ensemble de règles. Ces règles sont souvent établies par des experts en s'appuyant essentiellement sur des descriptions linguistiques, des dictionnaires de langue et des mots déclencheurs.

La deuxième approche est basée sur l'apprentissage automatique (orientée données), en utilisant des algorithmes d'apprentissage majoritairement supervisés sur un corpus de textes préalablement annoté, pour apprendre des règles d'extraction de manière autonome (par exemple, les champs aléatoires conditionnels (McCallum and Li, 2003; Settles, 2004), la machine à vecteurs de support (Vapnik, 1964; Kazama et al., 2002; Isozaki and Kazawa, 2002)[78 ; 79], le modèle de Markov caché (Bikel et al., 1998), etc.), ou non supervisés tels que le *clustering* (Etzioni et al., 2005; Nadeau et al., 2006). Il existe dans la littérature des systèmes de reconnaissance d'entités nommées qui combinent les deux techniques précédentes afin d'en tirer la maximum d'avantages. Pour cela, les règles d'extraction d'entités nommées sont soit apprises automatiquement puis révisées manuellement, soit écrites manuellement puis corrigées ou améliorées avec l'apprentissage automatique (Lin et al., 1998).

En résumé, le processus de fouille de texte et le traitement automatique de langage naturel en général, connaît plusieurs techniques de pré-traitement. Le choix de ces techniques a rarement été soutenu par la recherche empirique ou la théorie. De multiples travaux se basent sur des pré-traitements de texte « standards » tout en utilisant des techniques parfois contradictoires. En effet, chacune des techniques susmentionnées doit être étudiée et adaptée à la fois aux besoins et aux données concernées.

2.3 Classification automatique des données textuelles

Dans le cadre de cette thèse, nous avons travaillé essentiellement sur des données textuelles issues des centres de lutte contre le cancer en France. Ces données sont les comptes rendus médicaux écrits par les médecins souvent pendant ou après leur consultation des patients atteints du cancer. Ces comptes rendus sont non structurés et représentent une mine d'informations et une riche base de connaissances qui doit être impérativement exploitée pour guider les médecins lors de la prise de décision médicale, et ainsi améliorer le processus de santé. Cependant, le volume important de ces données rend leur exploration et leur analyse manuelle impossibles. Pour y remédier, il est tout naturellement logique de penser aux applications de fouille de texte et autres méthodes, puisqu'elles permettent d'extraire et de rechercher des informations pertinentes à partir de grands volumes de données textuelles.

La classification de textes est souvent mise en œuvre par des systèmes de traitement de l'information. Cette tâche permet de classer de façon automatique des données textuelles généralement en provenance d'un corpus dans une ou dans plusieurs catégories, en faisant appel à des méthodes numériques (c'est-à-dire des algorithmes de recherche d'information ou de classification de type mathématique). La classification permet d'organiser, de trier, et de rendre rapidement accessibles les informations contenues dans les corpus, et ainsi pouvoir en tirer le maximum d'avantages dans leur secteur d'activité, par exemple : la médecine.

La plupart des systèmes de classification de texte et de catégorisation de documents peuvent être décrits selon les quatre phases suivantes : (i) l'extraction de caractéristiques, (ii) la réduction de dimensions, (iii) la sélection de classificateurs, et (iv) l'évaluation des performances de techniques de classification. Dans un premier temps, nous allons voir les techniques de représentation de données textuelles et d'extraction de caractéristiques utiles.

2.3.1 Représentation des données et extraction des caractéristiques

Dans cette section, nous allons étudier de manière approfondie les approches de classification automatique de textes. L'application de ces dernières nécessitent que les données textuelles doivent être converties en « unités traitables ». En effet, la clé pour comprendre ces documents est leur représentation, qui servira comme entrée aux algorithmes de classification automatique de textes. Une solution intéressante permettant d'obtenir une représentation sémantique de documents textuels est de transformer chaque mot en un vecteur représentant sa sémantique. De nombreuses méthodes ont été utilisées pour déduire des vecteurs à partir de texte au niveau des caractères, des mots, des phrases ou des documents. Toutes ces méthodes visent à quantifier la richesse des informations et à les rendre plus adaptées pour l'apprentissage automatique, notamment la classification de textes.

L'extraction des caractéristiques d'un texte consiste à le représenter de manière appropriée pour une analyse sémantique nécessaire dans la plupart des tâches d'apprentissage automatique. Cette représentation nous permet de réaliser de nombreuses

statistiques utiles sur un ensemble de données, par exemple identifier les mots fréquents ou identifier le nombre de documents contenant un mot spécifique, etc.

De nombreux chercheurs ont proposé des techniques de représentation des données textuelles et d'extraction de caractéristiques. Une fois les données préparées, ces techniques peuvent être appliquées. Parmi les méthodes courantes d'extraction de caractéristiques : les modèles de sac de mots (*Bag of Words*), tels que la fréquence des termes (TF) (Salton and Buckley, 1988a), la fréquence inverse des documents (IDF) (Jones, 1972), Word2Vec (Mikolov et al., 2013) et les vecteurs globaux pour la représentation de mots (GloVe) (Pennington et al., 2014). L'extraction des caractéristiques de textes joue un rôle crucial dans la classification de textes puisqu'elle a un impact direct sur la précision de la classification de ces textes (Béchet et al., 2009).

2.3.1.1 TF ou Sac de mots (BoW)

La forme la plus élémentaire d'extraction des caractéristiques des mots pondérées est TF (*Term Frequency*), aussi appelée sac de mots (*Bag-of-Words*), où chaque mot est associé à son nombre d'occurrences dans l'ensemble des corpus. Les méthodes qui se basent sur TF utilisent généralement la fréquence des mots comme pondération booléenne ou logarithmique. Ainsi, chaque document est converti en un vecteur (de longueur égale à celle du document) contenant la fréquence des mots dans ce document. Bien que cette approche soit intuitive, elle est limitée par le fait que des mots couramment utilisés dans la langue peuvent dominer ces représentations. De plus, la relation sémantique entre les mots et l'ordre d'apparition dans la phrase ne sont pas sauvegardés.

2.3.1.2 TF-IDF

(Jones, 1972) a proposé la fréquence inverse des documents (IDF) comme méthode à utiliser en conjonction avec la fréquence des termes (TF) afin de réduire l'effet des mots couramment utilisés dans le corpus. La représentation mathématique du poids d'un terme dans un document par TF-IDF est donnée dans l'équation suivante :

$$W(d,t) = TF(d,t) * log(\frac{N}{df(t)}) \tag{2.1}$$

Ici N est le nombre de documents et $df(t)$ est le nombre de documents contenant le terme t dans le corpus. A cette méthode, l'Allocation de Dirichlet Latente (LDA) (Blei et al., 2003) ou l'Analyse Sémantique Latente (LSA) (Deerwester et al., 1990) sont souvent appliquées. Bien que cette mesure ait été proposée pour surmonter le problème des mots fréquemment utilisés, les relations sémantiques entre les mots, notamment les mots similaires, sont perdues puisque chaque mot est représenté indépendamment des autres. Cela engendre un réel problème quant à la compréhension des phrases d'un texte.

2.3.1.3 Plongement de mots (*Word embedding*)

Les représentations textuelles discrètes qui transforment les mots d'un ensemble de données en un espace de sac de mots causent une perte d'informations sémantiques.

Avec le développement de modèles plus complexes ces dernières années, de nouvelles méthodes ont vu le jour telles que le plongement de mots (*word embedding*) (Levy and Goldberg, 2014). Ces méthodes sont capables de surmonter ces limitations en représentant l'ensemble de données à l'aide d'une approche de sous-espace continu avec un nombre défini de composants ou de dimensions. Elles facilitent l'analyse sémantique et syntaxique car elles intègrent des concepts tels que le calcul de similarité entre les mots. Chaque mot du vocabulaire est représenté par un vecteur de N dimensions de nombres réels. Conceptuellement, c'est une intégration mathématique d'un espace multidimensionnel, où chaque dimension correspond à un mot, dans un espace vectoriel continu de dimension beaucoup plus faible. Ces représentations ont augmenté de manière significative les performances de diverses tâches de TALN telles que l'analyse syntaxique (Socher et al., 2013a) et l'analyse des sensations (Socher et al., 2013b). Word2Vec, GloVe et FastText (Joulin et al., 2016) sont les trois méthodes les plus courantes utilisées avec succès.

2.3.1.4 Word2Vec

Word2Vec est un groupe de modèles très puissants permettant de découvrir les relations entre les mots dans un corpus de données, notamment la similarité. Les mots dits similaires se voient attribuer des vecteurs proches dans l'espace vectoriel. L'architecture de ces modèles est composée d'un réseau de neurones artificiels formé de deux couches (une couche cachée), et est formé de manière à capturer le contexte sémantique des mots. Il s'agit des techniques les plus connues dans le TALN pendant ces dix dernières années. Les vecteurs de mots peuvent être calculés à l'aide de deux modèles : le modèle de sacs de mots continus (CBOW) et le modèle skip-gram. Dans les deux cas, la taille des plongements de mots est définie par la taille de la couche de projection.

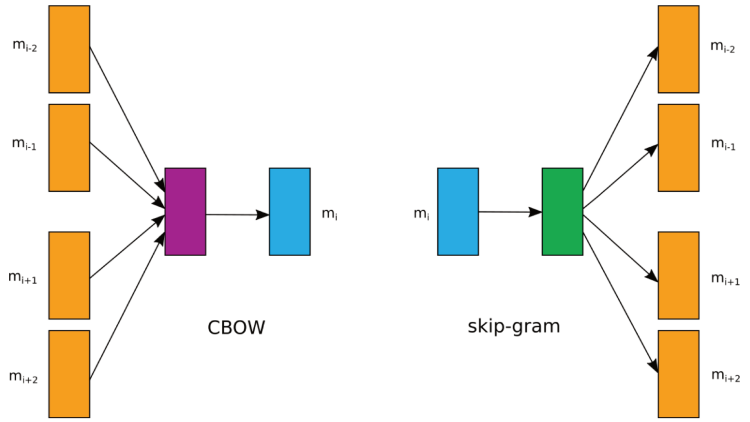


FIGURE 2.1 – Comparaison entre le modèle CBOW qui apprend une représentation (en violet) d'un mot m_i à partir des mots de son contexte ; et le modèle *Skip-gram* qui apprend une représentation (en vert) d'un mot, à partir de celui-ci pour en retrouver le contexte.

Le modèle CBOW permet d'obtenir une représentation appropriée d'un mot à partir des mots qui lui sont proches dans le texte. Tandis que le modèle skpi-gram vise

à prédire les mots du contexte étant donné un mot. D'après les auteurs de l'article (Mikolov et al., 2013), le modèle CBOW est rapide à entraîner, cependant, le modèle skip-gram fournit de meilleurs résultats pour les mots peu fréquents.

Les plongements de mots produits par les modèles CBOW et skip-gram peuvent être entraînés avec un très grand ensemble de données avec des millions de mots. Le modèle résultant peut être utilisé pour créer les vecteurs d'un ensemble de données cible. Cependant, certains mots du nouvel ensemble de données peuvent ne pas être connus du modèle entraîné. Une solution simple à cet inconvénient consiste à attribuer un vecteur fixe à tous les mots hors vocabulaire (*out-of-vocabulary*).

2.3.1.4.1 Glove

Une autre technique puissante pour le plongement de mots est GloVe (Pennington et al., 2014). Ce modèle est similaire à la méthode Word2Vec, et combine les avantages de la factorisation matricielle globale et des méthodes de contexte local. En effet, la formation est effectuée sur des statistiques globales de cooccurrence de mots du corpus. Le contexte est une fenêtre de longueur fixe d'éléments lexicaux centrés sur le mot. GloVe cherche à représenter chaque mot i et chaque mot j apparaissant dans le même contexte par des vecteur v_i et v_j respectivement, de dimension $d = [50, 100, 200, 300]$ tels que : $v_i.v_j + b_i + b_j = log(X_{ij})$ où X_{ij} représente le nombre de fois où le mot j se produit dans le contexte du mot i. b_i et b_j sont des biais scalaires associés aux mots i et j respectivement.

2.3.1.4.2 FastText

FastText (Joulin et al., 2016) est une bibliothèque logicielle pour l'apprentissage de plongements de mots et la classification de textes proposée par le laboratoire de recherche sur l'IA de Facebook. FastText est basé sur le modèle skip-gram de Word2Vec. Ce modèle permet d'apprendre les représentations de mots de manière à optimiser les performances de la classification d'un texte. La différence principale est que chaque mot est représenté par un sac de tous les n-grammes de caractères possibles qu'il contient. De cette manière, la morphologie du mot est donc prise en compte lors du calcul du vecteur, même si l'ordre des grammes est ignoré. Par exemple, si n = 3, le mot « traitement » donnera le sac de tri-grammes suivant : [<tr, tra, rai, ait, ite, tme, men, ent, nt> , <traitement>]. Le vecteur prend désormais en compte chaque n-gramme de caractères et le vecteur du mot est la somme de tous les vecteurs n-grammes de caractères du mot. Ainsi, FastText premet de résoudre la principale limitation de Word2Vec, à savoir les mots qui n'apparaissent pas dans le vocabulaire ne peuvent pas être représentés par le modèle.

Dans le modèle original de skip-gram, étant donné le corpus d'entraînement $W = \{w_1, w_2, ..., w_T\}$, l'objectif est de maximiser la fonction de vraisemblance logarithmique sur la base de la probabilité :

$$\sum_{t=1}^{T} \sum_{c \in C_t} log\ p(w_c|w_t) \tag{2.2}$$

Où C_t fait référence à l'ensemble des mots de contexte entourant le mot w_t. La manière la plus simple de définir la probabilité du mot de contexte w_c étant donné w_t est de calculer la fonction softmax définie comme suit :

$$softmax(s(w_c|w_t)) = \frac{e^{s(w_t,w_c)}}{\sum_{j=1}^{T} e^{s(w_t,w_j)}} \tag{2.3}$$

Où $s(w_c|w_t) = u_{w_t}^{\intercal}.v_{w_c}$. Les vecteurs $u_{w_t}^{\intercal}$ et v_{w_c} correspondent respectivement aux plongements des mots w_t et w_c.

Dans le modèle *FastText*, une fonction de score s différente est proposée. Étant donné un mot w_t composé par l'ensemble des n-grammes G_{w_t}, une représentation vectorielle z_{w_t} est associée à chaque n-gramme $g \in G_{w_t}$. Ainsi, la fonction de score est définie comme suit :

$$s(w_t, w_c) = \sum_{g \in G_{w_t}} z_g^{\intercal}.v_{w_c} \tag{2.4}$$

Cette fonction permet de partager les représentations entre les mots, et d'apprendre la représentation des mots rares.

2.3.1.5 Recommandations pour le calcul des vecteurs de mots

Il existe plusieurs plongements de mots entraînés à partir des corpus de données très larges, comme wikipedia, utilisant entre autres les méthodes mentionnées ci-dessus. Ces vecteurs sont disponibles sur internet, facilement accessibles et utiles pour avoir un modèle de représentation très général. Cependant, cela pourrait biaiser la modélisation et rendre moins précis le plongement de mots vis à vis d'un problème particulier. En d'autres termes, l'utilisation de tels modèles pourrait dissiper des différences entre vecteurs utiles pour une problématique donnée. L'idéal serait d'entraîner son propre plongement de mots à partir des données du corpus sur lequel on travaille quand on a le temps et les ressources nécessaires en termes de données.

2.3.2 Réduction de dimensionnalité

Les séquences de texte dans les modèles vectoriels basés sur des termes comportent de nombreuses caractéristiques. Cela rend la complexité du temps et la consommation de mémoire très coûteuses pour ces méthodes. Pour résoudre ce problème, de nombreux chercheurs utilisent la réduction de dimensionnalité pour réduire la taille de l'espace des caractéristiques. Ces méthodes permettent de projeter des données issues d'un espace de grande dimension dans un autre espace d'une dimension plus petite. De cette façon, elles réduisent la complexité des problèmes d'apprentissage automatique, notamment la classification de texte, et elles permettent d'améliorer des propriétés de stabilité et de robustesse de ces algorithmes (Bousquet and Elisseeff, 2002).

L'analyse de composantes principales (Jolliffe, 1986) est la techniques la plus répandue dans l'analyse multivariée et la réduction de la dimensionnalité. Elle consiste à

identifier un sous-espace dans lequel se trouvent approximativement les données passées en entrée (Abdi and Williams, 2010). PCA essaie de trouver de nouvelles variables non corrélées et maximiser la variance pour « préserver autant que possible la variabilité » (Jolliffe and Cadima, 2016). En dépit des éventuels avantages de l'utilisation de ces méthodes dans les tâches de traitement automatique de langage, cette étape reste facultative dans le processus de classification de textes.

2.3.3 Classification automatique des textes

Dans cette section, nous présentons quelques algorithmes d'apprentissage automatique utilisés dans la classification de texte parmi ceux qui sont les plus connus dans la littérature. Dans ce qui suit, nous abordons certains algorithmes d'apprentissage traditionnels tels que :*Naïve Bayes* (NB) (Edwards, 1986), les arbres de décision (DT) (Morgan and Sonquist, 1963), les machines à vecteurs de support (SVM) (Vapnik, 1964) et les k plus proches voisins (kNN) (Patrick and II, 1969). Ces algorithmes sont souvent rapides et précis quand ils sont appliqués au texte. Nous décrivons également des algorithmes basés sur des réseaux de neurones profonds tels que les réseaux de neurones convolutionnels (CNN) (LeCun et al., 1989) et les réseaux de neurones récurrents (RNN) (Rumelhart et al., 1985), ainsi que des techniques de combinaisons. Dans un premier temps, nous commençons d'abord par définir quelques notions de base de l'apprentissage automatique.

2.3.3.1 Apprentissage automatique

L'apprentissage automatique, dans le domaine de l'intelligence artificielle, permet aux ordinateurs d'apprendre à partir des données. Ce processus est généralement réparti sur trois phases. La phase d'entraînement qui consiste à former un modèle à partir de l'ensemble des données d'entraînement disponibles en nombre fini. La deuxième est celle de l'évaluation pendant laquelle le modèle entraîné est testé. La dernière phase correspond à la mise en production, c'est-à-dire de nouvelles données sont utilisées pour obtenir un résultat correspondant à une tâche donnée. Il arrive que certains systèmes continuent leur apprentissage une fois en production. Il existe plusieurs familles d'algorithmes d'apprentissage automatique selon les informations disponibles lors de la première phase et la tâche qu'on souhaite résoudre : apprentissage supervisé, apprentissage non supervisé, apprentissage semi-supervisé, apprentissage par renforcement, et apprentissage par transfert. Les deux premières familles sont aujourd'hui omniprésentes dans les applications de fouille de texte. Dans nos travaux de thèse, nous nous sommes intéressés particulièrement à ces deux familles d'algorithmes.

2.3.3.1.1 Apprentissage non supervisé

L'apprentissage non supervisé consiste à déduire des modèles sous-jacents à partir d'un ensemble de données non étiquetées, sans aucune référence à des résultats étiquetés. Il existe plusieurs algorithmes d'apprentissage non supervisé, cependant, le partitionnement des données (*clustering*) est de loin la technique la plus souvent utilisée. Son objectif est d'identifier automatiquement des sous ensembles à partir des données, où chacun regroupe des données ayant des caractéristiques similaires. Les

groupes de données obtenus, appelés «clusters», sont ensuite triés en fonction de leur pertinence. La notion de similarité se traduit par une distance D entre les données.

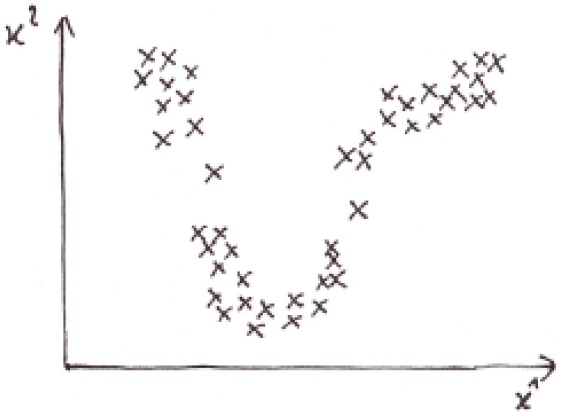


FIGURE 2.2 – Partitionnement des données.

Dans la Figure 2.2 ci-dessus, il semble naturel d'identifier trois regroupements. Notons qu'ils ne sont pas nécessairement isotropes ; ils n'ont pas le même nombre d'éléments, et parfois l'appartenance de certains points à un groupe est ambiguë.

Par ailleurs, d'autres algorithmes d'apprentissage non supervisé existent tels que les méthodes de réduction de dimension, par exemple, l'analyse en composantes principales (PCA) ou l'estimation de densités de probabilité. L'indexation sémantique latente (LSI) est aussi un autre type d'apprentissage non supervisé. Cette technique identifie les mots et les phrases qui se produisent fréquemment dans le même contexte.

Compte tenu de l'absence d'étiquettes, il est impossible aux algorithmes d'apprentissage non supervisé de calculer de façon sûre un score de réussite.

2.3.3.1.2 Apprentissage supervisé

Les algorithmes d'apprentissage automatique s'appliquent à des données étiquetées (c'est-à-dire annotées) qui constituent l'ensemble d'apprentissage $E = (x_n, y_n)$ de taille $N, 1 \leq n \leq N$. Pour obtenir des résultats satisfaisants, E doit être représentatif d'une population d'échantillons large. L'objectif de ces algorithmes est donc de définir une fonction de prédiction f_w (aussi appelée hypothèse) à partir des données annotées à des fins de généralisation sur des données non vues lors de la phase d'apprentissage. Ainsi, pour chaque nouvelle donnée $x \notin E$, f_w prédit une étiquette y associée à x à partir des connaissances fournies par les N exemples de l'ensemble d'apprentissage. On considère que la fonction de prédiction f_w appartient à une famille de fonctions paramétrées par w, où w représente un ensemble de paramètres, potentiellement très grand. La prédiction de l'étiquette de x sera alors $f_w(x)$. Lors de l'apprentissage, w est adapté de manière à optimiser les performances de prédiction sur l'ensemble d'apprentissage. De manière générale, l'algorithme mesure l'écart entre les étiquettes connues y_n des données de l'ensemble d'apprentissage, et les étiquettes prédites $f_w(x_n)$. Cependant, la minimisation de cet écart ne suffit pas pour obtenir des performances de prédiction optimales. Alors, on calcule l'étiquette d'une nouvelle entrée x : $f_{\widetilde{w}}(x)$, avec l'ensemble

des paramètres $\widetilde{w}$ obtenu lors de l'apprentissage. Le choix du critère d'optimisation de w détermine le type d'apprentissage supervisé : problème de régression ou de classification. Dans la régression, y_n sont des valeurs scalaires ou vectorielles. Tandis que dans la classification, y_n constitue un ensemble fini dont les éléments correspondent à des catégories ou des classes à identifier.

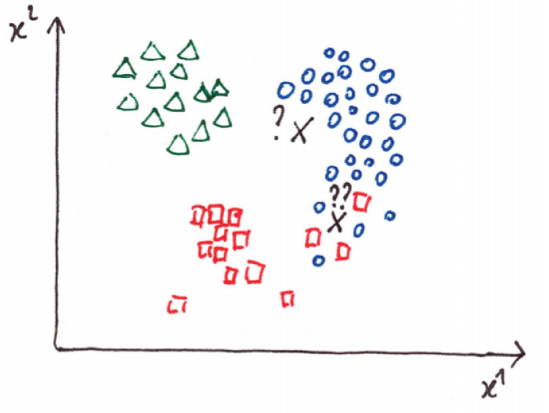


FIGURE 2.3 – Exemple de classification supervisée.

Dans la Figure 2.3 ci-dessus, les données sont annotées selon leur forme (triangle, rond, carré) en trois catégories. L'objectif est de déterminer la catégorie d'une nouvelle données non étiquetée. La classification supervisée permet de définir donc une partition de l'ensemble des données où chaque élément est attribué à l'une des classes.

2.3.3.2 Algorithmes de classification supervisée de texte

La classification supervisée peut s'appliquer à de nombreuses tâches dans des domaines d'applications différents tels que : la détection de fraudes (Irofti et al., 2020), le tri automatique de spams (courriers indésirables) (Cormack and Lynam, 2007), ou de documents (Larkey, 1999), la reconnaissance d'images (Zhang et al., 2020), etc., et à des données de nature différente par exemple : des textes, des images, des vidéos, etc. Dans le cadre de cette thèse, nous nous intéressons plus particulièrement à la classification supervisée de textes issus du domaine médical.

La classification de textes supervisée a pour objectif d'attribuer une ou plusieurs classes (catégories) à chaque texte (un document, un paragraphe, une phrase ou un mot) en fonction de son contenu sémantique et syntaxique. Il s'agit d'une tâche fondamentale du TALN avec de vastes applications dans le monde réel. Cette tâche nécessite des documents annotés afin d'apprendre les modèles de classification. Ensuite, ces modèles pourront classer les nouveaux documents dans la ou les classes appropriées. Parmi les algorithmes des plus connus, nous allons cité certains dans ce qui suit.

2.3.3.2.1 Classification naive bayésienne (NB)

Naive Bayes est un algorithme d'apprentissage génératif. Il s'agit de la méthode la plus traditionnelle de catégorisation de textes, et qui a été utilisée depuis les années

1950 (Porter, 1980). La méthode de classification Naïve Bayes fait partie d'une famille de classificateurs probabilistes basés sur l'application du théorème de Bayes (John and Langley, 1995), qui est formulé de la manière suivante :

— Soient $A_1, A_2, ..., A_L$ des événements mutuellement exclusifs dont l'union a une probabilité égale à 1. C'est-à-dire :

$$\sum_{i=1}^{L} P(A_i) = 1 \tag{2.5}$$

— On considère que les probabilités $P(A_i)$ sont connues.
— Soit B un événement tel que la probabilité conditionnelle de B sachant A_i, c'est-à-dire $P(B|A_i)$, est connue pour tout événement A_i.

Alors :

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} \tag{2.6}$$

Dans un problème de classification, $A_1, A_2, ..., A_L$ correspondent aux classes $C_1, C_2, ..., C_L$, et B correspond à un vecteur de caractéristiques X d'événements ($X_1 = x_1, X_2 = x_2, ..., X_m = x_m$), avec X_i qui représentent les variables et x_i qui représentent les valeurs. Alors, on peut réécrire la formule de Bayes (2.6) de la manière suivante :

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{2.7}$$

Dès lors que le dénominateur $P(X)$ est identique pour toutes les classes C_i, l'estimation de la probabilité sera exprimée comme suit :

$$P(C_i|X) = P(X|C_i)P(C_i) \tag{2.8}$$

Une fois les probabilités estimées, il s'agit désormais de classifier chaque nouvelle instance X en identifiant sa classe la plus probable selon la fonction :

$$f(X) = argmax P(Ci|X), i \in [1, L] \tag{2.9}$$

Cet algorithme est dit naïf car il repose sur l'hypothèse forte que les caractéristiques sont conditionnellement indépendantes les unes des autres, c'est-à-dire qu'elles présentent une indépendance conditionnelle de classe. En pratique, cette hypothèse n'est généralement pas respectée. Néanmoins, ce classificateur est utilisé dans diverses tâches de classification (Cheng et al., 2005; Muda et al., 2016), et est capable de rivaliser avec des méthodes de classification plus sophistiquées, en raison de son efficacité et sa rapidité.

2.3.3.2.2 Arbres de décision (DT)

L'arbre de décision (Morgan and Sonquist, 1963) est un algorithme d'apprentissage supervisé basé sur l'utilisation d'un arbre comme modèle de prédication. Les arbres de décision peuvent être utilisés pour la classification et pour la régression. Dans ces structures, les feuilles représentent les étiquettes de classe et les branches représentent les conjonctions des caractéristiques qui mènent à ces étiquettes de classe (c'est-à-dire les règles d'induction qui divisent l'espace d'instance en deux ou plusieurs sous-espaces). Les algorithmes de construction d'arbres de décision fonctionnent généralement de haut en bas, en choisissant à chaque étape une variable qui divise au mieux l'ensemble des éléments. La Figure 2.4 ci-dessous représente un exemple d'arbre de décision.

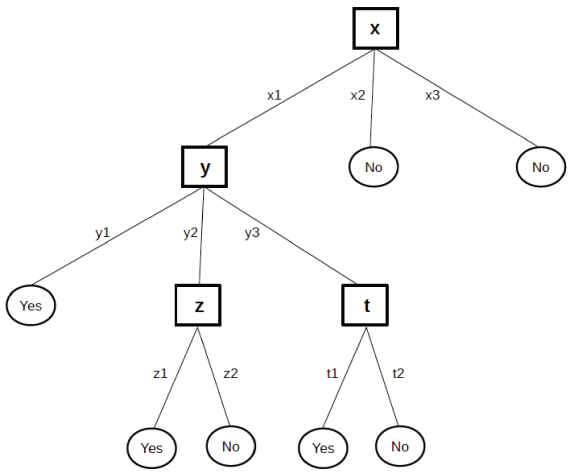


FIGURE 2.4 – Arbre de décision.

Pour construire un tel arbre, l'algorithme effectue une recherche gloutonne pour trouver à chaque itération la caractéristique qui divise le mieux l'espace d'instances en fonction de certaines métriques qui mesurent généralement l'impureté (ou la diversité) des variables cibles au sein des sous-ensembles. Le but est donc de trouver la division qui minimise l'impureté. Plusieurs méthodes ont été proposées pour trouver l'attribut optimal comme l'IG (Quinlan, 1986), l'indice de Gini (Breiman et al., 2017) ou l'entropie croisée.

IG est une métrique basée sur l'entropie de la théorie de l'information. Elle mesure la quantité d'information qu'une variable aléatoire apporte sur une autre variable aléatoire, c'est-à-dire la réduction de l'entropie d'une variable donnée par rapport à une autre. Étant donné un ensemble de données D avec n valeurs cibles (des classes), l'entropie de D est définie comme suit :

$$H(D) = -\sum_{i=1}^{n} p_i log_2(p_i) \tag{2.10}$$

Où p_i est la proportion d'éléments appartenant à la $i$ème classe. En termes d'entropie, $IG(D, V)$ d'une variable V par rapport à l'ensemble de données D est :

$$IG(D,V) = H(D) - H(D|V) \tag{2.11}$$

Avec :

$$H(D|V) = \sum_{v \in valeurs(V)} \frac{|D_v|}{|D|} H(D_v) \tag{2.12}$$

Ici, Dv est un sous-ensemble de D où V a la valeur v.

Contrairement à l'entropie et à l'indice de Gini (Lerman and Yitzhaki, 1984), IG mesure la pureté (Lee and Lee, 2006). Ainsi, la caractéristique qui maximise l'IG est utilisée pour diviser l'espace de l'instance.En revanche, l'indice de Gini mesure l'hétérogénéité d'un ensemble de données, c'est une mesure très similaire à l'entropie. Étant donné un ensemble de données D avec N classes, et p_i la proportion d'éléments appartenant à la $i$ème classe, l'indice de Gini I de l'ensemble de données est défini par :

$$I(D) = 1 - \sum_{i=1}^{N} p_i^2 \tag{2.13}$$

2.3.3.2.3 Machines à vecteurs de support (SVM)

Les machines à vecteurs de support ont été développées par (Vapnik, 1964), et font partie des algorithmes de classification supervisée les plus populaires dans l'apprentissage automatique. L'objectif des SVM est de regrouper les éléments de différentes classes dans différentes zones de l'espace vectoriel d'entrée en utilisant un hyperplan de marge maximale. Cet hyperplan est considéré comme un séparateur optimal avec une meilleure capacité à classifier de nouvelles données inconnues. Dans le modèle SVM, les instances sont représentées sous forme de points (vecteur réel) dans cet espace.

Étant donné une tâche de classification binaire où l'étiquette d'instance est représentée par $y_i = 1$ si x_i appartient à la classe positive et $y_i = 0$ si elle appartient à la classe négative, et un ensemble de données d'apprentissage de n instances représentées comme suit :

$$(\overrightarrow{x}_1, y_1), (\overrightarrow{x}_2, y_2), ..., (\overrightarrow{x}_n, y_n)$$

Dans l'espace vectoriel, tout hyperplan pourrait être exprimé comme suit :

$$\overrightarrow{w}.\overrightarrow{x} - b = 0 \tag{2.14}$$

Dans le cas où les données d'apprentissage sont linéairement séparables, SVM tente de trouver un hyperplan de marge maximale qui sépare dans l'exemple de la Figure 2.5 les petits ronds des petits carrés en résolvant le problème d'optimisation suivant :

$$Minimiser\ ||\overrightarrow{w}||\ ; y_i(\overrightarrow{w}.\overrightarrow{x_i} - b) \geq 1,\ \ for\ i \in [1,n]$$

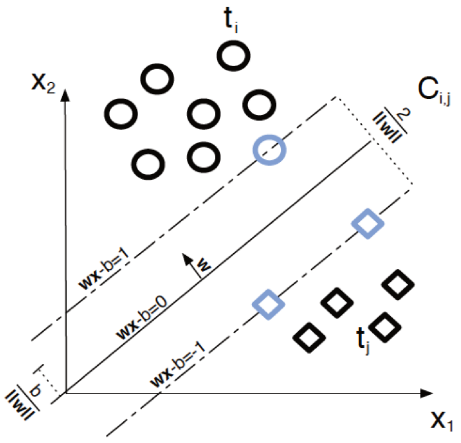


FIGURE 2.5 – Exemple de classification par SVM.

Enfin, la fonction de classification est donc la fonction «signe» suivante :

$$y = f(\overrightarrow{w}) = sgn(\overrightarrow{w}.\overrightarrow{x_i} - b) \tag{2.15}$$

Où $\overrightarrow{x}$ et b sont les solutions au problème d'optimisation ci-dessus. Pour les données d'apprentissage non linéairement séparables, l'hyperplan qui produit le moins de prédictions de classification incorrectes sera sélectionné. Cela se fait par le biais de la perte de charnière l (*hinge loss*) (Rosasco et al., 2004) :

$$l(y) = max(0, 1 - y_i.y) \tag{2.16}$$

On peut constater à partir de la formule 2.16 ci-dessus que la perte de charnière est égale à 0 lorsqu'on prédit la bonne classe, et $|y| \geq 1$. Cependant, une pénalité est appliquée lorsque $|y|$ est inférieur à 1 ou lorsque la prédiction est erronée (y et y_i ont un signe opposé). Cette pénalité augmente linéairement avec y. Le problème d'optimisation à minimiser devient :

$$\left[\frac{1}{n}\sum_{i=1}^{n} max(0, y_i(\overrightarrow{w}.\overrightarrow{x_i} - b))\right] + \lambda||\overrightarrow{w}||^2 \tag{2.17}$$

Une autre alternative, quand les données d'apprentissage ne sont pas linéairement séparables, est de projeter les vecteurs des données dans un espace de plus grande dimension de manière à ce qu'ils soient linéairement séparables. Ensuite, l'algorithme SVM est utilisé pour trouver l'hyperplan optimal qui sépare ces nouveaux vecteurs de données.

2.3.3.2.4 Les k plus proches voisins (kNN)

L'algorithme des k plus proches voisins (Patrick and II, 1969) fait partie des algorithmes dits «paresseux». Il n'effectue pas un apprentissage proprement dit, mais

pour chaque nouvelle instance de test x, l'algorithme kNN trouve les k voisins les plus proches de x parmi toutes les instances de l'ensemble d'apprentissage, et note les candidats de catégorie en fonction de la classe de k voisins. L'hypothèse sur laquelle s'appuie l'algorithme kNN consiste en l'idée qu'une instance est plus proche des instances de la même classe que celles des autres classes. Par conséquent, lors de la classification d'une instance inconnue, l'algorithme regarde la classe la plus nombreuse parmi les classes des k plus proches instances. La Figure 2.6 ci-dessous, montre un exemple d'architecture d'un modèle kNN simplifié (ensemble de données en 2D).

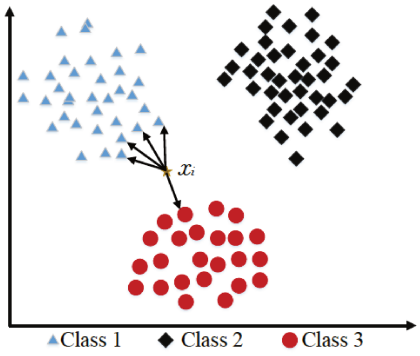


FIGURE 2.6 – Une architecture de modèle kNN pour l'ensemble de données 2D et trois classes.

La règle de décision de kNN est :

$$f(x) = arg\ \max_{j}\ S(x, C_j) = \sum_{d_{i \in KNN}} sim(x, d_i) y(d_i, C_j) \tag{2.18}$$

où S fait référence à la valeur de score par rapport à $S(x, C_j)$, la valeur de score du candidat i à la classe de j, et la sortie de $f(x)$ est une étiquette vers l'instance de l'ensemble de tests.

Le choix de la valeur de k a un effet sur la performance du classificateur kNN. En effet, un k très petit pourrait rendre le système sensible au bruit, en revanche, le choix d'un k trop grand pourrait provoquer un sous-apprentissage.

2.3.3.3 Apprentissage automatique profond

Les modèles d'apprentissage profond ont obtenu des résultats de pointe dans de nombreux domaines, y compris une grande variété d'applications TALN. L'apprentissage profond pour la classification de textes comprend trois architectures de base. Dans ce qui suit, nous allons décrire chacun de ces modèles.

2.3.3.3.1 Réseau de neurones profonds

Les réseaux de neurones profonds (*Deep Neural Network*) sont conçus de manière à apprendre à partir d'une multi-connexion de couches, où chaque couche reçoit uniquement la connexion de la couche précédente, et fournit des connexions uniquement à

la couche suivante (Kowsari et al., 2017). La Figure 2.7 ci-dessous illustre la structure d'un DNN standard.

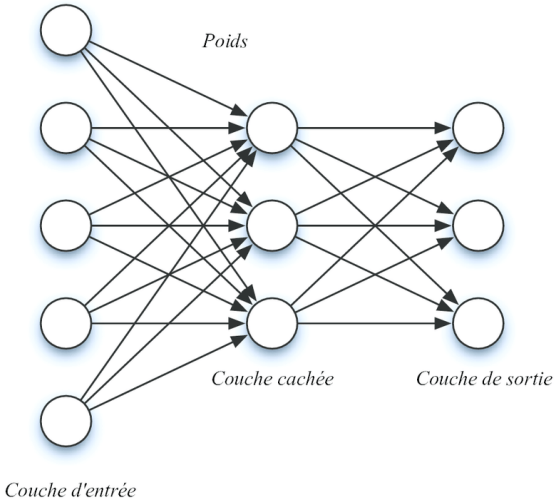


FIGURE 2.7 – Un perceptron multi-couches contenant 3 couches : couche d'entrée de 5 neurones, une couche cachée de 3 neurones et une couche de sortie de 3 neurones.

La couche d'entrée peut être construite avec les techniques d'extraction de caractéristiques, telles que TF-IDF ou une méthode de plongement de mots. La couche de sortie contient autant de neurones que de classes disponibles pour la tâche de classification (un seul neurone pour une classification binaire). Dans les DNN multi-classes, chaque modèle d'apprentissage est généré (le nombre de nœuds dans chaque couche et le nombre de couches sont attribués). DNN est un modèle discriminatif qui utilise un algorithme de rétro-propagation standard utilisant sigmoïde (2.19), ReLU (Nair and Hinton, 2010) (2.20) comme fonction d'activation, et une fonction Softmax (2.21) pour la classification multi-classes.

$$f(x) = \frac{1}{1 + e^{-x}} \in (0, 1) \tag{2.19}$$

$$f(x) = max(0, x) \tag{2.20}$$

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \forall j \in \{1, ..., K\} \tag{2.21}$$

Étant donné un ensemble de paires (x, y), $x \in X$, $y \in Y$, le but est d'apprendre la relation entre les données d'entrée et les données cibles en utilisant les couches cachées.

2.3.3.3.2 Réseau de neurones récurrents (RNN)

Une autre architecture de réseau de neurones souvent utilisée dans les applications de fouille et de classification de textes est le réseau de neurones récurrents (*Recurrent*

Neural Network) (Sutskever et al., 2011; Mandic and Chambers, 2001). RNN attribue plus de poids aux points de données précédents d'une séquence. Ce qui fait de cette technique une méthode puissante et adaptée à la classification de textes et aux données séquentielles de manière générale (Graves et al., 2013). Un RNN considère les informations des nœuds précédents dans une méthode sophistiquée qui permet une meilleure analyse sémantique de la structure d'un ensemble de données. Dans la classification de texte, RNN fonctionne principalement en utilisant LSTM (Hochreiter and Schmidhuber, 1997) ou GRU (Chung et al., 2014), comme indiqué sur la Figure 2.8 ci-dessous.

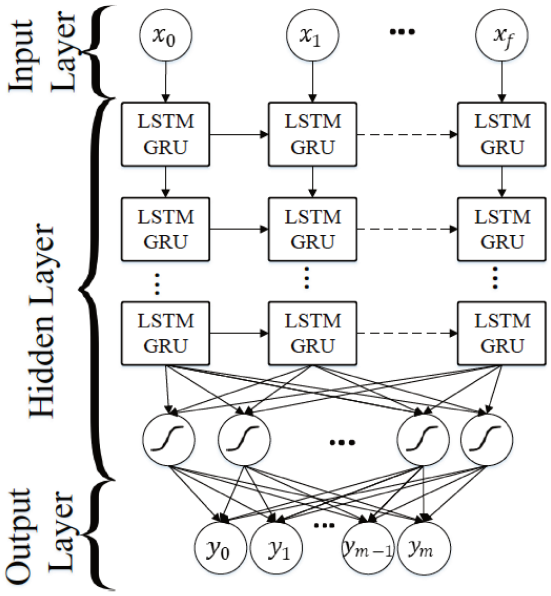


FIGURE 2.8 – Réseau de neurones récurrent standard (LSTM/GRU).

Cette méthode peut être formulée de la manière suivante :

$$x_t = F(x_{t-1}, u_t, \theta) \tag{2.22}$$

où x_t est l'état du réseau à l'instant t et u_t représente l'entrée à l'étape t. Concrètement, on utilise des poids pour formuler l'équation (2.22), paramétrée par :

$$x_t = W_{rec}\sigma(x_{t-1}) + W_{in}u_t + b \tag{2.23}$$

où W_{rec} désigne le poids récurrent de la matrice, W_{in} se réfère aux poids d'entrée, b est le biais et σ désigne une fonction élément par élément.

Malgré les bonnes performances des RNN appliqués aux tâches de classification de textes, ils restent vulnérables aux problèmes d'explosion et de disparition de gradient lorsque l'erreur de descente de gradient se propage à travers le réseau (Bengio et al., 1994).

2.3.3.3.3 Long Short-Term Memory (LSTM)

Le LSTM a été introduit par (Hochreiter and Schmidhuber, 1997)), et a depuis été enrichi par de nombreux chercheurs (Graves and Schmidhuber, 2005). Il s'agit d'un type spécial de RNN qui résout les problèmes d'explosion et de disparition de gradient (Pascanu et al., 2013) en préservant la dépendance à long terme d'une manière plus efficace par rapport au RNN de base. Bien que LSTM ait une structure en forme de chaîne similaire à RNN, LSTM utilise plusieurs portes pour réguler soigneusement la quantité d'informations qui est autorisée dans chaque état de nœud. La Figure 2.9 ci-dessous montre la cellule de base d'un modèle LSTM.

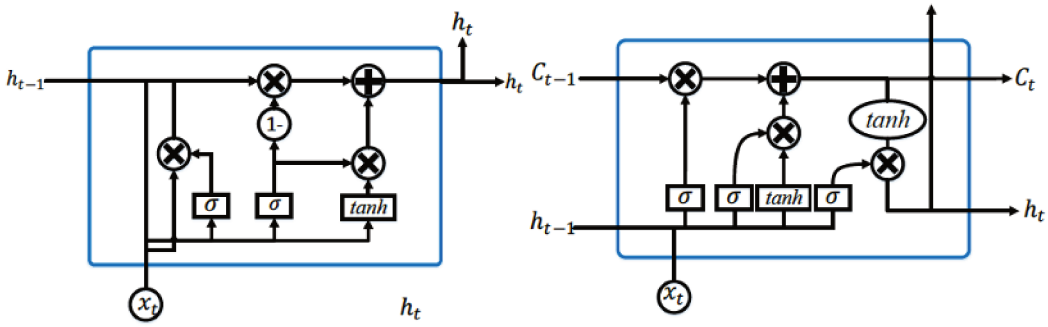


FIGURE 2.9 – À gauche c'est une cellule GRU, et à droite c'est une cellule LSTM

.

Une explication étape par étape d'une cellule LSTM est la suivante :

$$i_t = \sigma(W_i\,[x_t, h_{t-1}] + b_i) \tag{2.24}$$

$$\widetilde{C_t} = tanh(W_c\,[x_t, h_{t-1}] + b_c]) \tag{2.25}$$

$$f_t = \sigma(W_f\,[x_t, h_{t-1}] + b_f) \tag{2.26}$$

$$C_t = i_t * \widetilde{C_t} + f_t C_{t-1} \tag{2.27}$$

$$o_t = \sigma(W_o\,[x_t, h_{t-1}] + b_o) \tag{2.28}$$

$$h_t = o_t tanh(C_t) \tag{2.29}$$

où l'équation (2.24) représente la porte d'entrée, l'équation (2.25) représente la valeur de la cellule mémoire candide, l'équation (2.26) définit l'activation de la porte d'oubli, l'équation (2.27) calcule la nouvelle valeur de la cellule mémoire et les équations (2.28) et (2.29) définissent la valeur finale de la porte de sortie. Dans la description ci-dessus, chaque b représente un vecteur de biais, et chaque W représente une matrice de poids, et x_t représente l'entrée dans la cellule de mémoire au temps t. Les indices i, c, f, o désignent respectivement les portes d'entrée, de mémoire de cellule, d'oubli et de sortie. La Figure 2.9 ci-dessus montre une représentation graphique de la structure de ces portes.

2.3.3.3.4 Unité récurrente fermée (GRU)

Les GRU sont un mécanisme de déclenchement pour réseaux de neurones récurrents formulé par (Chung et al., 2014; Cho et al., 2014). Les GRU sont une variante simplifiée de l'architecture LSTM. Cependant, une GRU diffère du LSTM car elle contient deux portes et ne possède pas de mémoire interne (C_{t-1} sur la Figure 2.9). De plus, une deuxième non-linéarité n'est pas appliquée ($tanh$ sur la 2.9). Une explication étape par étape d'une cellule GRU est la suivante :

$$z_t = \sigma_g(W_z x_t + U_z h_{t-1} + b_z) \tag{2.30}$$

où z_t fait référence au vecteur de portes de mise à jour de t, x_t représente le vecteur d'entrée, W, U et b représentent les matrices/vecteurs de paramètres. La fonction d'activation σ_g est soit sigmoïde soit ReLU et peut être formulée comme suit :

$$\widetilde{r}_t = \sigma_g(W_r x_t + U_r h_{t-1} + b_r) \tag{2.31}$$

où r_t représente le vecteur de portes de ré-initialisation de t, z_t est le vecteur de portes de mise à jour de t.

$$h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \sigma_h(W_h x_t + U_h(r_t \circ h_{t-1}) + b_h) \tag{2.32}$$

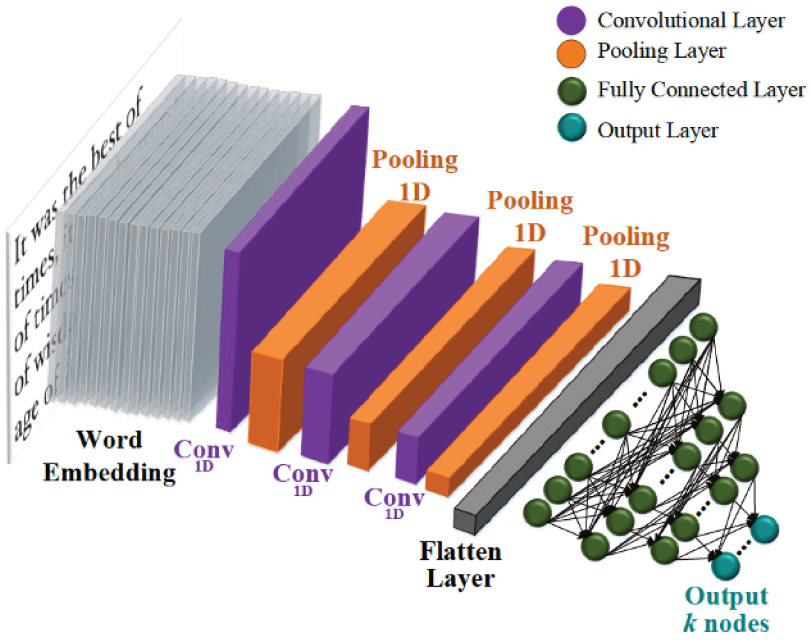
où h_t est le vecteur de sortie de t, et σ_h indique la fonction tangente hyperbolique.

Les réseaux de neurones récurrents peuvent être biaisés lorsque dans un texte les mots ultérieurs sont plus influents que les précédents. Des modèles de réseaux de neurones convolutionnels (CNN) ont été utilisés afin de surmonter ce biais en déployant une couche de *pooling* max pour déterminer les phrases discriminantes dans les données textuelles (Lai et al., 2015).

2.3.3.3.5 Réseau de neurones convolutionnel (CNN)

Un réseau neuronal convolutif (CNN) est une architecture d'apprentissage profond couramment utilisée pour la classification hiérarchique des documents (Lai et al., 2015; Jaderberg et al., 2016). Bien qu'à l'origine conçus pour le traitement d'image, les CNN ont également été utilisés de manière performante dans la classification de textes (LeCun et al., 2015; Lecun et al., 1998). Dans un CNN de base pour le traitement d'images, un noyau de convolution $d * d$ va analyser les caractéristiques de l'image en entrée. Ces couches de convolution peuvent être empilées pour fournir plusieurs filtres sur l'entrée. Pour réduire la complexité de calcul, les CNN utilisent des couches de *pooling* «mise en commun» pour réduire la taille de la sortie d'une couche à l'autre dans le réseau. Différentes techniques de mise en commun sont utilisées pour réduire les sorties tout en préservant les caractéristiques importantes (Scherer et al., 2010). La méthode la plus connue est *max pooling*, où l'élément maximal de la fenêtre de regroupement est sélectionné. Afin d'alimenter la sortie groupée des cartes empilées vers la couche suivante. Les couches finales d'un CNN sont généralement entièrement connectées.

En général, pendant l'étape de rétro-propagation d'un CNN, les poids et les filtres du détecteur de caractéristiques sont ajustés. Un problème potentiel qui se pose lors de l'utilisation de CNN pour la classification de textes est le nombre de «canaux» S (taille de l'espace des fonctionnalités). Alors que les applications de classification d'images ont généralement peu de canaux (par exemple, seulement 3 canaux RVB), S peut être très grand (par exemple, $50K$) dans la classification de texte (Johnson and Zhang, 2014), ce qui est traduit par une dimensionnalité très élevée.



7

FIGURE 2.10 – Architecture d'un réseau de neurones convolutionnels (CNN) pour la classification de texte.

La Figure 2.10 ci-dessus illustre l'architecture d'un CNN pour la classification de textes composé d'une couche d'entrée $1D$ remplie avec des vecteurs extraits via une méthode de plongement de mots, des couches de *pooling* $1D$, des couches entièrement connectées et enfin une couche de sortie (Kowsari et al., 2019).

2.3.4 Évaluation des techniques de classification

La dernière partie d'un système de classification de textes consiste en son évaluation. Dans la recherche, il est primordial d'avoir des mesures de performance comparables afin d'évaluer les algorithmes. Cependant, de telles mesures n'existent que pour quelques de méthodes. Lors de l'évaluation des méthodes de classification des textes, l'absence de protocoles de collecte de données standard représente un réel défi. De plus, choisir des ensembles d'entraînement et de test différents peut introduire des incohérences dans le calcul des performances d'un modèle. Un autre défi concernant l'évaluation des méthodes de classification de textes est de pouvoir comparer différentes mesures de performance utilisées dans des expériences séparées. En effet, les mesures de performance évaluent généralement des aspects spécifiques de la performance des tâches de classification et ne présentent donc pas toujours des informations identiques.

La compréhension des performances d'un modèle est essentielle à l'utilisation et au développement des méthodes de classification des textes. Dans la littérature, il existe de nombreuses méthodes pour évaluer les techniques supervisées. Le F-score est l'un des paramètres d'évaluation agrégés les plus populaires pour l'évaluation des classificateurs (Hossin and Sulaiman, 2015; Liu et al., 2014; Lever et al., 2016).

2.4 Modélisation de la résistance aux traitements

Malgré les progrès de l'amélioration de la prise en charge du cancer et l'augmentation significative de l'arsenal thérapeutique, les phénomènes d'échappement et de résistance aux traitements restent des problématiques conduisant aux décès de nombreux patients chaque année. La résistance peut être caractérisée soit de *de novo*, et dans ce cas, aucun effet significatif sur le volume tumoral n'est observé après la première ligne de traitement ; soit d'acquise pour les cas où après une réponse plus ou moins longue à la première ligne de traitement, une reprise de croissance de la tumeur primaire, voire une dissémination métastatique de la tumeur est observée (Hazlehurst and Dalton, 2006).

La résistance *de novo* est une incompatibilité d'emblée entre la première ligne de traitement et la tumeur. L'amélioration de la connaissance de la tumeur et de son environnement devrait augmenter l'adéquation entre la première ligne de traitement et le phénotype tumorale et ainsi diminuer le risque d'apparition de résistance *de novo* (Syn et al., 2017).

La résistance acquise, quant à elle, repose sur des mécanismes spécifiques de la thérapie utilisée. Les années 1990 ont vu l'arrivée et le succès clinique des thérapies ciblées ainsi que l'apparition de phénomènes de résistance associée à ces thérapies. La connaissance précise du mécanisme d'action de ces thérapies a permis une identification rapide de certains acteurs de la résistance à ces thérapies (Paulson et al., 2018).

La vision mono-paramétrique du phénomène de résistance centrée sur la cellule tumorale a été abandonnée au profit d'une vision multi-paramétrique et donc nettement plus complexe du phénomène de résistance dont l'origine peut être la cellule tumorale mais également son micro-environnement. La cellule tumorale n'est donc plus considérée comme la seule cible possible pour l'action thérapeutique. L'émergence des thérapies immuno-modulatrices complexifie un peu plus l'appréhension des phénomènes de résistance. En effet, les nombreux efforts engagés pour améliorer et augmenter le niveau de caractérisation de la pathologie tumorale adressent principalement la composante tumorale de la tumeur et ne renseignent pas ou peu sur le microenvironnement tumoral. Les techniques d'immunohistologie restent à ce jour les seules à apporter de l'information dans ce domaine, au niveau clinique.

2.4.1 Travaux connexes

Plus nous en savons sur le cancer, plus sa complexité devient évidente. L'analyse d'une pathologie spécifique nécessite souvent d'identifier un nombre important de patients qui transcendent un niveau d'institut unique. De nombreuses recherches

fondamentales ont été réalisées afin de mieux comprendre les mécanismes impliqués dans le développement du cancer de manière générale (Ducreux, 2019; Chanchou et al., 2020; Duhamel et al., 2020). Toutefois, il persiste encore beaucoup d'interrogations sur cette maladie.

De nouvelles méthodes d'exploration et de compréhension sont donc nécessaires à développer afin d'aider les spécialistes dans leur travail, notamment pour la compréhension des différentes interactions intervenant dans le développement des tumeurs. Parmi le peu de travaux qui se sont focalisés sur la modélisation de la résistance aux traitements anticancéreux ou la réponse aux traitements de manière générale, nous pouvons citer les travaux suivants (Colin et al., 2014; Cornelis et al., 2013).

Les progrès qu'a connu ce domaine sont en grande partie dûs aux avancées réalisées dans l'imagerie médicale. Les techniques d'imagerie permettent aux médecins de vérifier la présence d'une tumeur, sa taille, sa forme, son activité métabolique et sa localisation exacte. Ces informations sont utiles pour définir les traitements à mettre en œuvre et pour évaluer si la chirurgie peut être proposée. Dans ces études les chercheurs proposent des modèles mathématiques décrivant l'évolution (ou la croissance) des métastases, où le calcul de l'efficacité d'un traitement est basée essentiellement sur la mesure de diamètre de la plus large lésion.

La quasi totalité de ces travaux dans la littérature se base uniquement ou essentiellement sur les données d'imagerie (CT-scan, IRM, échographie, PET-scan). Or, plus de 80% des informations cliniques pertinentes se trouvent dans le texte des dossiers patient (données textuelles).

Dans le chapitre 6 de ce mémoire, nous présenterons une approche originale pour l'identification et la modélisation des résistances aux traitements.

CHAPITRE 3

PROJET CONSORE

Sommaire

Dans ce chapitre nous allons voir quelques détails techniques du projet Consore dans lequel s'inscrit cette thèse. Compte tenu de l'importance et la complexité du projet, ce chapitre nous permet de mettre l'accent sur les éléments les plus importants du projet et qui intéressent notre thèse tels que l'architecture globale du projet, les technologies utilisées, les données sur lesquelles nous avons travaillé et les modèles utilisés pour définir les concepts métiers.

3.1 Introduction

Défini par une multitude de pathologies distinctes, le cancer tue environ 150000 personnes chaque année en France. L'analyse d'une pathologie spécifique nécessite souvent d'identifier un nombre important de patients qui se situent souvent au-delà d'un seul centre. Jusqu'à présent, l'identification de ces patients était une tâche chronophage qui nécessite d'énormes ressources et une interprétation manuelle des dossiers patients. Dans ce contexte, la fédération nationale des centres de lutte contre le cancer UNICANCER (regroupant 18 centres) — l'une des plus grandes fédérations européennes spécialisées dans la recherche et la prise en charge du cancer — a lancé le projet ConSoRe (**Con**tinuum **So**in **Re**cherche) en partenariat avec l'entreprise Sword. ConSoRe est un outil innovant qui accompagne les médecins et chercheurs quotidiennement dans leur travail. Ce projet ConSoRe est déployé dans 9 centres de cancérologie et leur offre un puissant moteur de recherche permettant de fouiller dans les dossiers de santé électroniques (DSE) afin d'identifier rapidement des cohortes de patients selon les critères choisis par les médecins. Le projet aborde quatre préoccupations majeures : (i) L'agrégation d'une énorme quantité de données hétérogènes en temps réel ; (ii) L'analyse sémantique des dossiers de santé électroniques, standardisation des données et modélisation de la maladie cancéreuse ; (iii) La mise en œuvre technique d'une solution permettant une interrogation rapide des données à l'échelle nationale ; (iv) La création de services prêts à l'emploi pour les cliniciens et les chercheurs.

3.2 Motivations et projets similaires

UNICANCER a lancé en 1975 une enquête nationale («Enquête Permanente Cancer», EPC (Colombani et al., 2013)) pour collecter des données structurées auprès de patients cancéreux recherchant un traitement dans l'un des centres de la fédération. Une enquête riche en informations médicales telles que la localisation de la tumeur, le code histologique, le classement de la tumeur et les traitements des patients, mais aussi des données administratives, telles que des données démographiques (nom, prénom, date de naissance ou adresse). Un effort entrepris initialement par les services d'information médicale, il a progressivement diminué au fur et à mesure que l'équipe d'origine était dispersée dans des postes non spécialisés.

En 2010, UNICANCER a décidé de mettre fin au projet, suite à la recommandation du conseil d'administration, il s'est engagé dans la création d'une infrastructure nationale permettant aux médecins et aux chercheurs cliniciens d'accéder directement aux données de santé stockées dans le système d'information de chaque centre.

Suite à un examen exhaustif de l'état de l'art de l'entreposage de données cliniques, l'initiative caBIG, (Warden, 2011), a été identifiée comme une plate-forme de recherche collaborative appropriée. Cependant, sa modularité et sa dépendance à des normes communes pour garantir l'interopérabilité des données présentent des réels inconvénients. Un autre système largement utilisé est l'outil i2b2, (Murphy et al., 2010), des Centres nationaux de calcul biomédical (NCBC) aux États-Unis. Ce système offre une architecture puissante pour stocker et interroger des données structurées, mais avec une exploration et une visualisation médiocres des données. De plus, le système n'est pas intuitif pour les médecins et ne peut pas gérer les rapports médicaux textuels non structurés. Ce dernier point est commun à d'autres plate-formes de recherche translationnelle [1] qu'on a identifiées telles que tranSMART, (Canuel et al., 2015).

En ce qui concerne les outils basés sur le traitement automatique de langage naturel, cTAKES, (Savova et al., 2010), est l'un des systèmes performants souvent utilisé dans l'analyse et l'extraction sémantique. Mais celui-ci ne convient pas car il est adapté au traitement des textes rédigés en anglais, ce qui ne répond pas à notre problématique. En parallèle de nos travaux, d'autres initiatives françaises d'implémentation d'entrepôts de données cliniques ont vu le jour comme eHOP, (Madec et al., 2019), ou Dr. Warehouse, (Ivančević et al., 2013), qui est plus orienté vers les rapports narratifs. Ces deux outils se concentrent sur la médecine générale ou les maladies rares et offrent des fonctionnalités avancées pour l'exploration des données de santé, mais n'ont pas été conçus spécifiquement pour la recherche sur le cancer.

3.3 Architecture du projet

ConSoRe a une infrastructure évolutive capable de gérer des collections de données hétérogènes allant de 2 à 16 millions de documents sources dans les centres où il est déployé. Une solution capable de stocker les concepts extraits, les données inférées et les relations existantes entre toutes les données du modèle. De plus, des exigences strictes en matière d'interrogation : des réponses rapides (< 2 secondes) et reproductibles ont impacté le choix de l'architecture du modèle de données. L'architecture du projet est décrite par le schéma de la Figure 3.1.

Compte tenu des considérations cliniques et des cas d'utilisation répertoriés par les médecins, plusieurs sources de données pertinentes on été intégrées dans ConSoRe : la descriptions du patient, les comptes rendus cliniques, les fiches d'administration de chimiothérapie, le groupe homogène de diagnostic (DRG), les caractéristiques tumorales (issues de l'enquête EPC), des échantillons biologiques et des essais cliniques. Ces données représentent des millions de documents de nature hétérogène : SQL, XML, CSV, MS Word, texte, PDF ou documents numérisés. Les documents traités sont parfois structurés (tels que : PMSI, CRB, Fiche tumeur, etc.), mais pour la plus grande partie, ils n'ont aucune structure définie, en particulier les comptes rendus. Pour assurer l'interopérabilité, tous les formats de fichiers sont basés sur des standards existants

1. La recherche translationnelle se situe entre la recherche fondamentale, dont le travail consiste à comprendre les mécanismes à l'origine du développement d'un cancer, et la recherche clinique qui vise à évaluer l'efficacité et la tolérance de nouveaux traitements sur les patients.

et largement utilisés (PAMS (Jongkind et al., 1993), (Slavov et al., 2013), PN13/XML).

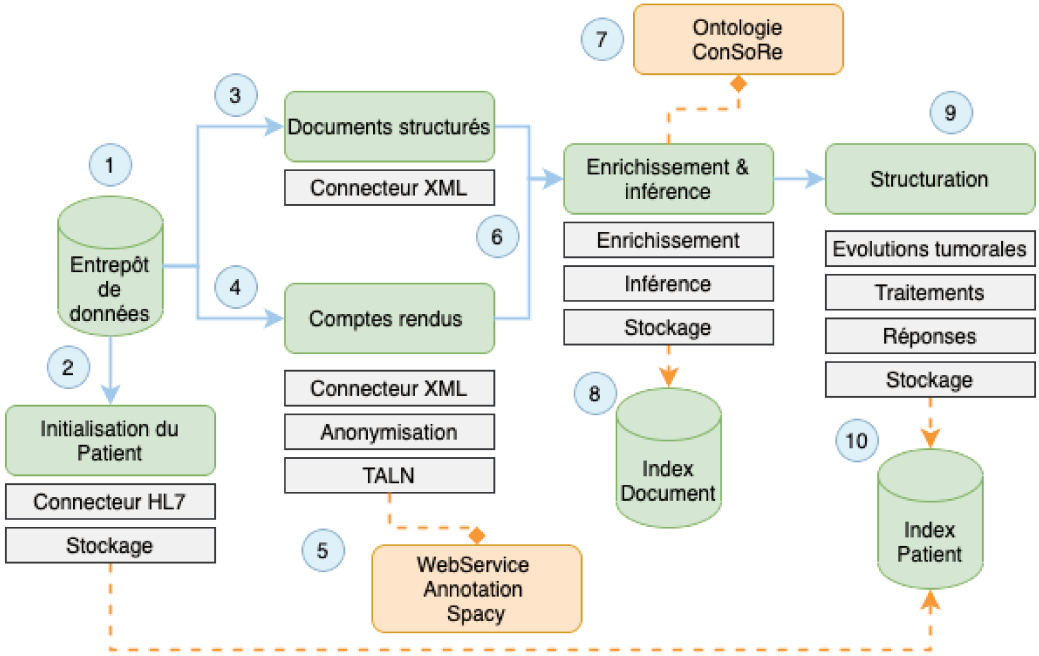


FIGURE 3.1 – Chaîne de traitement de Consore.

Une chaîne de traitement récupère un ensemble des fichiers à traiter. Ceux qui contiennent des données non structurées (les comptes rendus) sont envoyés à un service de traitement automatique du langage naturel pour extraire les concepts à partir des textes. Ces comptes rendus peuvent être de plusieurs types : prescriptions médicamenteuses, compte rendus d'imagerie (IRM, Scanner, . . .), comptes rendus de consultation, de fin de séjour hospitalier, etc. Tous ces documents n'ont pas la même valeur puisque les informations, qu'ils contiennent, n'ont pas le même degré de fiabilité. Par exemple, la suspicion de présence de tumeur dans un compte rendu d'imagerie ne contient pas la même quantité d'information qu'un compte rendu indiquant clairement un diagnostic ou le début d'un traitement de chimiothérapie. De plus, les comptes rendus ne proviennent pas des mêmes centres de cancérologie. Ce qui rend la définition d'un format pivot, dans lequel on stocke les informations, une tâche très complexe puisque chaque médecin a son propre style d'écriture et chaque centre a ses propres conventions.

L'ensemble des concepts, extraits à partir des documents, sont enrichis à l'aide de différents référentiels (cf. Annexe A). Pour cela, plusieurs normes nationales ou internationales ont été adoptées : CIM-10 (Classification Internationale des Maladies, version 10) pour les codes de diagnostic, SNOMED 3.5VF qui contient un thésaurus de 115 000 termes cliniques (en particulier, les codes de topographie et de morphologie), l'ADICAP est un thésaurus des codes de pathologie utilisés dans les centres de cancérologie français, CCAM («Classification Commune des Actes Médicaux») est une nomenclature française pour coder les actes médicaux, Vidal est un thésaurus français

des médicaments largement utilisés dans les applications de prescription médicale du pays. Vidal fournit un tableau d'équivalence sur les codes ATC (*Anatomical Therapeutic Chemical*), MIABIS 2.0 est une norme européenne qui vise à normaliser les éléments de données utilisés pour décrire les biobanques, la recherche sur les échantillons et les données associées et est soutenue par le consortium BBMRI-ERIC.

Une fois enrichi, le résultat est enregistré dans l'index *Document*. Ensuite, tous les documents enrichis sont agrégés et l'historique du patient est structuré à l'aide d'un système expert composé de plusieurs règles classées en deux catégories : celles qui sont impliquées dans la détection d'événements, et celles qui sont responsables du calcul des événements des propriétés. Le dossier du patient est reconstitué, tous les actes réalisés sont détectés ainsi que les différents traitements de chimiothérapie. Une fois structuré, le patient est enregistré dans l'index *Patient*.

Dans l'index *Document*, on y stocke les données afférentes au document ainsi que les concepts élémentaires qui en sont extraits. En revanche, l'index *Patient* est rempli à deux moments différents de la chaîne de traitement. D'abord, au lancement de la chaîne de traitement à l'aide des fichiers de type état civil pour enregistrer les informations signalétiques du patient comme le nom, la date de naissance, etc. Ensuite, après la structuration où le patient est mis à jour avec l'ajout des concepts inférés à l'aide des documents lors de la structuration.

Enfin, ConSoRe est muni d'un portail web, sur lequel il est possible d'effectuer des recherches sur les patients en fonction de différents critères, tels que la date de début d'un cancer, la localisation d'un cancer, les traitements médicaux effectués, etc.

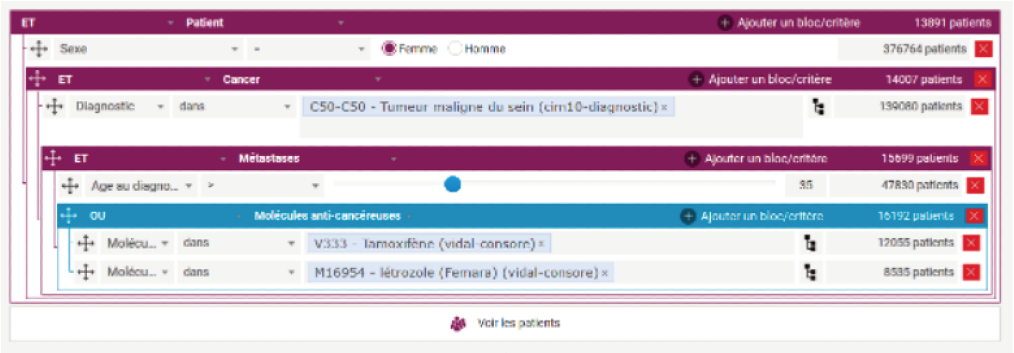


FIGURE 3.2 – Générateur de requêtes du Portail web de ConSoRe.

Le cœur de ConSoRe repose donc sur des pipelines de traitement de données de nature hétérogène, qui gèrent la création d'un aperçu complet de l'historique d'un patient à partir de centaines de fichiers XML fournis par chaque centre de cancérologie.

La frise chronologique indique les différents évènements qui sont arrivés au patient au fil du temps. C'est un bon moyen pour visualiser rapidement le dossier du patient. Le tableau de bord permet de suivre l'évolution du traitement et d'évaluer la qualité

des résultats une fois le traitement terminé. Le portail de recherche peut être utilisé pour rechercher et visualiser les documents analysés et les patients structurés. La Figure 3.3 montre une frise créée pour un patient.

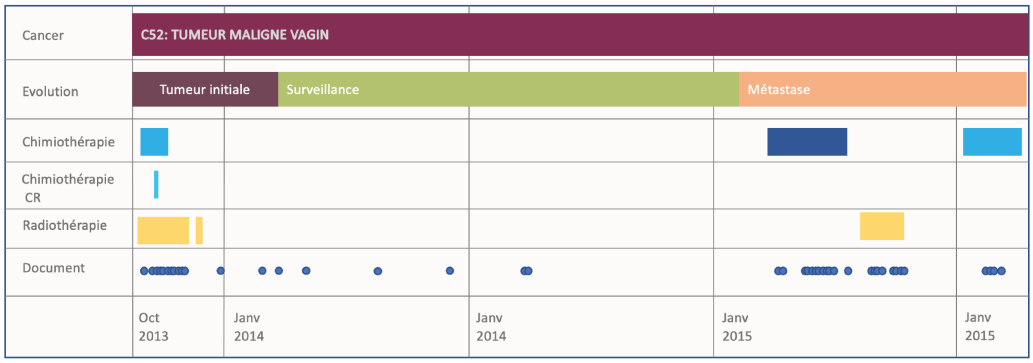


FIGURE 3.3 – Exemple d'une frise chronologique d'un patient.

3.4 Données

Les données manipulées dans ce projet sont très variables et proviennent de sources différentes : fiches de chimiothérapie, fiches tumeur, PMSI, CRB, comptes rendus, etc. Les données appartiennent à plusieurs centres de cancérologie spécialisés dans des pathologies différentes et représentatives des mécanismes de résistances et d'insensibilité en oncologie : il s'agit du cancer du sein, cancer du pancréas et cancer du poumon. Les documents structurés correspondent à des formulaires qui indiquent par exemple les actes médicaux effectués, les médicaments utilisés, les traitements chimiothérapiques. Les comptes rendus représentent la source de données non structurées et constituent plus de 80% de l'ensemble des données du projet.

Ces rapports sont de types très variés tels que : des comptes rendus d'imagerie (IRM, Scanner, ...etc), comptes rendus de première consultation, prescriptions médicamenteuses, etc. On y trouve plusieurs informations à caractère médical comme les décisions cliniques des médecins, les traitements administrés aux patients, les antécédents (qui regroupent entre autres les interventions chirurgicales antérieures, les hospitalisations et les maladies chroniques chez les membres de la famille), et les conclusions de fin de consultation ; mais aussi d'autres informations administratives comme celles relatives au centre (nom de l'établissement, adresse, numéro de téléphone, etc.), noms des médecins présents, etc.

De manière générale, dans le domaine de l'oncologie, le formatage de texte libre est couramment utilisé pour la rédaction des comptes rendus médicaux. En dépit des efforts consacrés dans quelques centres pour la saisie des données cliniques dans un format structuré, la plupart de ces rapports sont maintenus en textes libres, et contiennent des jargons spécifiques et ne suivent généralement pas de règles grammaticales formelles.

Ils sont souvent rédigés dans un style axé sur les mots clés et formatés en utilisant de nombreux sauts de ligne, espaces blancs, des listes à puces ou de dénombrement, etc. Des erreurs d'orthographe sont également présentes (inversion de deux caractères, suppression des apostrophes, langage abréviatif, etc.), et produisent des non-mots absents du lexique et donc difficilement analysables. De plus, l'écriture des médecins n'est pas toujours facilement liée aux termes standards (par exemple : le code CIM-10 «C50 / néoplasme malin du sein» est généralement associé à un large éventail de termes du «cancer du sein» à « carcinome canalaire invasif»).

De plus, comme chaque centre dispose de son propre système d'information et maintient ses données sous un ou plusieurs formats (PDF, MS Word, HTML ou fichier texte), le traitement de ces derniers peut engendrer des défis supplémentaires comme le traitement de données "sales". Par exemple, les phrases extraites d'un PDF *"océrisé"* sont parfois mal découpées ou représentées par une séquence insignifiante de caractères spéciaux entraînant du bruit.

Plusieurs difficultés sont identifiées dans ce projet. D'une part la validité des extractions faites à partir des documents non structurés. D'autre part, la difficulté est de pouvoir déterminer, à partir des données extraites des documents, la date de début d'une pathologie cancéreuse (tumeur, récidive, métastase...) ainsi que sa localisation. En effet, on peut par exemple extraire d'un document la chaîne de caractères « cancer du sein » sans garantir à 100% que le patient soit porteur d'une tumeur.

En raison des différences intrinsèques des styles de formatage qui varient selon les médecins et les institutions, nous avons mis en annexe (cf. Annexe B) quelques comptes rendus afin d'illustrer quelques variations des informations qu'ils peuvent contenir.

CHAPITRE 4

PRÉ-TRAITEMENT

Sommaire

4.1 Introduction

La phase de pré-traitement dépend en grande partie de la qualité et de la nature des textes traités. Si ces derniers sont normalisés et écrits dans une langue uniforme, alors cette tâche peut comporter des opérations linguistiques standards telles que la *tokensation*, la *segmentation en phrases*, la *lemmatisation*, etc. En revanche, s'ils sont de qualité moindre (divergences typographiques, erreurs d'orthographe ou de syntaxe, etc), il faut prévoir d'autres tâches de pré-traitement comme le nettoyage des données (détection et suppression du bruit) ou la correction orthographique.

Dans ce chapitre, nous allons présenter les méthodes de pré-traitement de texte développées et appliquées à des comptes rendus médicaux. Ces rapports sont une mine d'informations et constituent une base riche de connaissances qui doit être exploitée afin d'améliorer le processus de soins de santé. Cependant, compte tenu de l'absence d'une structure définie pour ces textes (cf. Annexe B), leur exploitation devient une tâche compliquée. Il est dès lors nécessaire de pré-traiter les textes bruts afin de pouvoir les exploiter à posteriori avec des processus unifiés et non une multitude de processus adaptés à tous les cas possibles. Pour ce faire, différentes techniques de préparation des données ont été mises en œuvre pour rendre facilement accessibles ces informations. Ces pré-traitements sont une étape cruciale constituée de toute sorte de transformations effectuées sur le texte pour garantir des données préparées pour les tâches ultérieures, telles que la classification automatique des textes, qui seront abordées dans le chapitre suivant 5.

Nous allons présenter dans ce qui suit les différentes techniques utilisées dans le cadre de notre recherche et présenter les expérimentations effectuées ainsi que les résultats obtenus.

4.2 Collecte, nettoyage et transformation des données

En conformité avec le règlement général sur la protection des données (RGPD), chaque centre membre de la fédération UNICANCER dispose de son propre système d'informations (SI) avec une infrastructure du projet *ConSoRe* comme système distribué avec un entrepôt de données local. Chaque entrepôt contient les données de quelques centaines de milliers de patients avec un nombre important de documents textuels de nature hétérogène, pouvant aller jusqu'à 16 Millions de documents pour un seul centre.

Dans le cadre de nos travaux, un entrepôt de données a été constitué avec de centaines de milliers de documents, issus des différents centres, afin de constituer un corpus de données représentatif, compte tenu de la particularité des données de chaque centre, et des considérations cliniques répertoriées par les médecins. L'ensemble de ces données a été anonymisé en respect avec les conditions établies par la CNIL (Commission Nationale de l'Informatique et des Libertés), la commission française de protection des données. Ce corpus intègre plusieurs sources de données pertinentes telles que les fiches tumeur, les fiches de chimiothérapies, les PMSI, les comptes rendus, etc. Les comptes

rendus constituent la source de données non structurées et représentent la grande partie des données. Plus de 360.000 comptes rendus ont été intégrés dans la création de ce corpus. Ces derniers connaissent une dizaine de types tels que les prescriptions médicamenteuses, les comptes rendus d'imagerie, les comptes rendus de consultation, etc. De plus, ces données sont stockées sous des formats différents qui dépendent parfois du centre duquel elles proviennent : XML, MS Word, document texte, PDF ou HTML.

Ces comptes rendus contiennent une variété d'informations médicales, telles que les décisions cliniques prises par les médecins, les traitements administrés aux patients, les antécédents (qui comprennent, entre autres, les chirurgies antérieures, les hospitalisations et les maladies chroniques des membres de la famille), mais également des informations administratives telles que le nom de l'établissement, les adresses, les numéros de téléphone, etc.

Le nettoyage de ce corpus doit permettre d'éliminer les bruits et les parties les moins utiles du texte qui, peuvent réduire les performances de tâches ultérieures, qui sont en réalité plus significatives, telles que la classification de phrases ou l'extraction d'entités nommées, tout en conservant les caractéristiques textuelles essentielles pour ces tâches. Pour ce faire, les traitements suivants ont été effectués sur l'ensemble des comptes rendus :

— Un filtre a été mis en place sur le nom du fichier pour le traiter une seule fois.
— Les documents PDF, XML, MS Word et HTML ont été parsés pour extraire le texte brut.
— Les documents HTML peuvent être encodés dans divers formats : ASCII, ISO 8859-1, Windows-1252, UTF-8, etc. Tous les formats ont été convertis en UTF-8 afin de faciliter l'utilisation ultérieure de ces comptes rendus.
— Les balises HTML des comptes rendus concernés ont été retirées. Les tableaux sont transformés en phrases. Certaines balises nécessitent un traitement spécifique pour assurer une bonne détection des limites de phrase (cf. 4.4). Les lignes de script et de style sont aussi retirées.
— L'identifiant des documents patients ont été anonymisés.
— Les suites de caractères vides dans une ligne ont été remplacées par un seul espace, et chaque succession de lignes vides (inutile) dans un document a été remplacée par une seule ligne vide.
— Les caractères spéciaux et les sauts de lignes abusifs ont été retirés afin de réduire la dimensionnalité du problème. Cela permet d'améliorer les performances de la tâche de classification. Il a été constaté que ces caractères sont fréquemment présents dans les documents PDF mal océrisés.
— Les caractères de ponctuation et les mots vides ont été gardés car ils sont sémantiquement pertinents pour l'analyse des rapports cliniques. Dans la phrase «*tumeur pulmonaire (cancer primitif ?)*», le médecin utilise le point d'interrogation afin de manifester un doute sur un éventuel cancer primitif.

4.3 Tokenisation

La tokenisation consiste à diviser un document texte en unités de traitement significatives appelés *tokens*. Un *token* est une séquence contiguë non vide de caractères, isolée de ses voisines par des séparateurs. Ces séparateurs sont souvent définis en fonction de la langue utilisée. En langue française, l'espace et les signes de ponctuation sont souvent utilisés comme séparateurs de mots. La tokenisation est une opération très importante dans le traitement automatique de langage naturel, et constitue une étape cruciale puisque toutes les tâches de pré-traitement de textes ultérieures se basent fondamentalement sur cette opération, notamment l'identification des limites de phrases, l'extraction d'entités nommées et la détection des dates. De manière générale, savoir où appliquer un découpage pour l'obtention des *tokens* n'est pas une tâche aisée. La nature morphologique des mots d'une langue influe grandement sur la sélection possible des points de découpage dans une chaîne textuelle.

L'intégralité des données utilisées dans le cadre de nos recherches est écrite en français. Le processus de tokenisation est donc basé sur un modèle spécifique en utilisant des règles et des expressions régulières. Cependant, afin de répondre de manière optimale à la tâche de tokenisation des comptes rendus, il est primordiale d'une part de prendre en compte la nature des informations véhiculées (le contenu et le style utilisés), et d'autre part, de garder à l'esprit les objectifs à atteindre derrière cette analyse de texte.

Dans le cadre de notre thèse, une phase d'analyse a d'abord été effectuée afin de consulter le maximum de comptes rendus possibles pour comprendre et répertorier les variantes des écritures utilisées dans le domaine de la cancérologie. Plusieurs constatations ont été faites. Les acronymes et les abréviations sont souvent utilisés par les médecins. Les URL et les adresses e-mail sont aussi présentes dans ces comptes rendus. Les mots techniques du domaine (les médicaments, les protocoles de chirurgie, les codes ADICAP (cf. Annexe A), les biomarqueurs, les noms des tumeurs, les codes SNOMED (cf. Annexe A), etc) constituent un jargon particulier qui échappe aux règles auxquelles pourrait inspirer une tokenisation standard. Plusieurs métriques sont utilisées afin d'évaluer par exemple l'évolution de la taille d'une tumeur, les doses des traitements administrés ou encore les résultats de différents examens cliniques. Certains de ces exemples issus des comptes rendus sont illustrés dans les phrases suivantes.

- `CA 15-3 à 117`
- `En échographie, la masse est mesurée à 39 mm`
- `Un nodule de perméation du QIE de 20mm`
- `Hypercalcémie modérée à 2.75 pour N < 2.6.`

D'après ces exemples, le point ne marque pas toujours la fin d'une phrase, mais il peut être contenu dans divers types de *tokens* comme les nombres, les dates ou encore les abréviations. Les termes contenant des traits d'union doivent impérativement être considérés comme une seule entité (un seul *token*). On note également la présence de termes médicaux tels que les biomarqueurs qui regroupent dans la même chaîne des caractères alphanumériques et des caractères spéciaux par exemple : «HER2 ++», «PD-L1», «PDGFRA/FGFR1».

En plus des règles strictes basées essentiellement sur les signes de ponctuation, tous les référentiels (cf. Annexe A) utilisés lors de l'enrichissement sémantique dans le cadre du projet ont été intégrés afin de bien identifier les mots du jargon médical. Cela a permis de définir et de prendre en compte les différentes exceptions de tokenisation.

Cette étape de tokenisation constitue non seulement une étape clé pour le restant des opérations de pré-traitement, mais aussi joue un rôle important dans la tâche de classification de phrase que nous allons voir dans le chapitre 5.

4.4 Détection des limites de phrases

4.4.1 Définition du problème

Afin de pouvoir utiliser les textes des comptes rendus, il est nécessaire de les soumettre à un certain nombre de traitements tels que le découpage des textes en phrases et la vérification de la langue que nous allons voir par la suite (cf. 4.5).

Une fois nettoyés, les textes sont présentés dans des fichiers sous forme de séquences de caractères : lettres, nombres, symboles de ponctuation ou autres. Afin de les traiter, il est important de pouvoir identifier les phrases. Ces dernières vont permettre de comprendre le sens des informations contenues dans les comptes rendus. Cette étape joue un rôle important et a un impact direct sur les résultats des tâches postérieures de notre chaîne de traitement automatique de langage naturel. C'est sur cette étape que repose les tâches suivantes notamment la classification des phrases et la reconnaissance d'entités nommées, mais aussi d'autres tâches de pré-traitement telles que l'identification de la portée des phrases négatives et incertaines (cf. 4.9) ou le rattachement de certains concepts identifiés, par exemple associer une localisation à une tumeur. Pour ce faire, un mécanisme de détection des limites de phrases (début et fin) a été mis en place pour transformer chaque compte rendu en une suite de phrases. Cette approche dépend en particulier des performances de la tâche de tokenisation (cf. 4.3) et s'appuie essentiellement sur les signes de ponctuation et la capitalisation des caractères.

4.4.2 Challenges et analyse de la solution

Pour tenir compte du lexique spécifique utilisé dans les rapports cliniques, comme pour la tokenisation, une étude a été menée afin d'identifier le maximum de cas variés possibles à gérer et choisir l'approche la plus appropriée à mettre en place. Plusieurs cas ont été identifiés. Quelques exemples issus des comptes rendus sont illustrés dans les phrases suivantes.

- Examen effectué le 12-09-96 (consult. le 08-11-96)
- Dr. XXXXXX J./SEC
- Examen :
  Abdomen tout à fait normal ; pas de masse ; pas d'hépatomégalie
- Au total
  2 problemes : une symptomatologie douloureuse dorsale haute et

un CA 15 3 élevé
- Bilan biologique du 03.06.2019 : ACE a 2.2 - CA 19-9 a 13
- cancer primitif identifié.
- Bruits du coeur reguliers, souffle systolique a 4/6eme au foyer aortique. pas de palpitation. pas de douleur
- Tabac :
- 1265 RIS DX2439795 XXXXXX brouillon 18/11/2019 16 :53 :00 18/11/2019 16 :53 :00 18/11/2019 16 :53 :00 Imagerie Paris SCANNER TH + AP AVEC INJECTION Scanner CR preliminaire Majoration de l'infiltration retroperitoneale perivasculaire autour de l'axe iliaque interne gauche

À partir de ces exemples, on peut constater que le point «.» et le retour à la ligne ne sont pas des marqueurs exclusivement utilisés dans une fin de phrase. Dans les comptes rendus, il est souvent question d'une seule et même phrase écrite sur plusieurs lignes. En d'autres termes, il existe des phrases contenant des retours à la ligne suivis de caractères alphanumériques. Les phrases ne commencent pas forcément par un retour à la ligne suivi par une majuscule ou un point suivi par une majuscule. Les comptes rendus issus de PDF mal océrisés peuvent contenir des suites de caractères longues et parfois incompréhensibles sans aucun marqueur de début ou de fin de phrase. Les informations qui y sont contenues sont sémantiquement très variées et peuvent regrouper des informations diverses. Cela complique davantage les choses lors de la structuration des comptes rendus. Certaines phrases peuvent paraître incomplètes, comme c'est le cas de l'avant dernière phrase dans les exemples présentés ci-dessus. L'absence de mot est parfois interprétée par les médecins. Dans cet exemple, le médecin veut transcrire le fait que le patient consulté n'est pas fumeur.

Plusieurs types de phrases sont alors identifiées : des phrases bien formées, des phrases incomplètes, et des phrases avec des parties disfluentes (contenant des pauses silencieuses ou des répétitions, etc.).

4.4.3 Approche proposée

Deux approches connues s'offrent à nous pour détecter les limites de phrases : soit en utilisant des règles (Wang and Huang, 2003), soit en entraînant un modèle d'apprentissage automatique (Rudrapal et al., 2015). Compte tenu de la complexité de la tâche et de la particularité des données manipulées, la détection des limites de phrases mise en place est basée sur des règles définies. Ce choix est justifié notamment par deux éléments. D'abord, lors de la phase d'analyse, il s'est avéré que les phrases incomplètes et les phrases ayant des parties disfluentes sont non représentatives de l'ensemble des phrases. De plus, les antécédents des patients n'ont souvent aucune caractéristique sémantique qui pourrait les dissocier des autres informations médicales contenues dans les rapports cliniques. Cependant, il s'agit souvent d'une succession de «phrases» constituant un seul bloc (ou paragraphe) souvent recopiées par les médecins au début d'un comptes rendu, comme illustré ci-dessous.

- 07/2013 :
  Douleurs aine gauche. Irradiation cotyle gauche. Progression ganglionnaire et osseuse : CARBOPLATINE-ALIMTA 4 cycles puis ALIMTA seul.

Afin de faciliter la tâche de classification automatique des phrases (cf. chapitre 5), Ces blocs de textes sont considérés comme une seule et même phrase. De cette manière, la date est assignée à toutes les parties de la phrase et pourrait constituer l'information sémantique qui caractérise les antécédents.

Plusieurs expérimentations ont été réalisées afin de valider cette approche. Pour cela, une vingtaine de comptes rendus ont d'abord été manuellement annotés. Ce qui est très peu si l'on considère la taille du corpus global ou le nombre de comptes rendus par centre. Cependant, il faut prendre en considération que ce travail de sélection et d'annotation s'est fait manuellement constituant une tâche laborieuse. Toutefois, les documents sélectionnés ont été choisis de manière à couvrir tous les types de comptes rendus possibles, et que les variantes qui y sont contenues sont les moins triviales à résoudre. Pour chaque document, l'ensemble des phrases attendues a été manuellement construit. L'algorithme parcourt alors les documents et les découpe en phrases. Une phrase est considérée comme bien découpée si et seulement si elle a les mêmes *tokens* de début et de fin que ceux de la phrase correspondante dans l'ensemble de test et si les deux phrases ont exactement le même nombre de *tokens* et le même numéro d'ordre. Sur un ensemble de 855 phrases, une exactitude de $96,4\%$ a été obtenue.

4.5 Correction orthographique

Comme évoqué dans les chapitres précédents et illustré dans les exemples en Annexe B, les comptes rendus sont sujets à des fautes d'orthographe et de syntaxe. Ces fautes sont de natures différentes et peuvent prendre plusieurs formes : effacement, insertion, substitution, permutation, coupure ou soudure.

Dans le but de mieux observer ce phénomène et d'appréhender la manière dont il peut impacter nos tâches ultérieures du traitement automatique de texte, plusieurs comptes rendus issus de notre corpus ont été consultés. Parmi les constatations établies, les fautes d'orthographe sont assez fréquentes. Cependant, leur correction n'est pas profitable dans le cadre de nos recherches. En effet, d'une part, cette tâche est très fastidieuse et chronophage au vu de ce qu'elle pourrait apporter comme avantages pour le traitement de texte. D'autre part, les plongements de mots utilisés pour la représentation des textes (cf. chapitre 5) permettent d'attribuer aux mots ayant une similarité syntaxique ou sémantique proche des vecteurs de mots proches dans l'espace vectoriel. Toutefois, un dictionnaire de langue française de plus de 336 530 mots a été utilisé afin de corriger les oublis d'accent.

4.6 Lemmatisation et Racinisation

C'est deux techniques de pré-traitement sont souvent utilisées à des fins de normalisation de texte. La lemmatisation effectue une analyse morphologique complète d'un mot et mappe ses formes flexionnelles à une forme de base commune. Par exemple, les mots « est » et « sont » sont regroupés sous un même lemme «être». Ainsi, la lemmatisation permet d'augmenter la couverture des dictionnaires de vocabulaire. Cependant,

appliquée à tout le corpus, cette tâche peut supprimer les différences de style. La racinisation est une technique qui associe un mot à sa racine. Elle est souvent basée sur des heuristiques utilisant des règles afin de supprimer les suffixes de mots, sans tenir compte des homographes. Cela pourrait produire des mots irréels (qui n'existent pas dans la langue utilisée) ; ce qui les rend difficiles à interpréter. De manière générale dans les tâches de TALN (Chowdhary, 2020), la lemmatisation peut être utile lorsqu'elle est appliquée au corpus pour augmenter la couverture des dictionnaires, en revanche, la racinisation devrait être appliquée à la fois au corpus et aux dictionnaires. Comme indiqué dans la section 2.2, certaines tâches de pré-traitement peuvent supprimer des informations utiles ou introduire des erreurs d'analyse qui réduisent la précision et la validité des tâches ultérieures.

Après une analyse effectuée sur les comptes rendus, une forme de lemmatisation a été appliquée sur les différents référentiels (dictionnaires), notamment les référentiels des actes chirurgicaux, des diagnostics et des évolutions tumorales.

4.7 Détection des dates

4.7.1 Définition du problème

Les expressions temporelles sont utilisées dans le langage naturel pour donner des informations sur le moment où quelque chose s'est passée, combien de temps ça a duré ou à quelle fréquence. Les dates et les heures sont souvent utilisées à ce propos. Dans le cadre de nos travaux, on s'intéresse particulièrement aux expressions temporelles permettant de localiser un événement dans le temps. Celles qui permettent de répondre à la question : *quand ?* Ces expressions prennent généralement plusieurs formes : des phrases nominales, des adverbes (tels que : « Aujourd'hui » ou « Hier ») ou de simples dates en utilisant soit des nombres, soit des lettres, soit les deux. La détection de ces éléments est importante dans le traitement automatique de textes, notamment lorsqu'il s'agit de traiter des textes cliniques.

En effet, cela permet de dater les informations extraites à partir de ces textes et ainsi, bien comprendre le sens de ces dernières. Dans nos recherches, la détection de ces informations contribuent essentiellement dans la structuration des patients via l'exploitation des documents patients, mais aussi dans la reconstitution du parcours du soins des patients. De plus, ces informations sont des expressions souvent utilisées dans la formulation des antécédents médicaux d'un patient. L'identification de ces antécédents permet d'aider les médecins dans leur travail notamment lors de la prise de décisions médicales en choisissant les traitements les plus appropriés.

4.7.2 Travaux connexes

Dans la littérature, plusieurs travaux se sont intéressés à la détection des événements temporels. Tapi Nzali et al. (2015) ont conçu une approche pour l'extraction automatique d'expressions temporelles. Ils montrent que la distribution des ces expressions est liée au domaine. Les autres soulignent que l'adaptation de certaines tâches

de pré-traitement, telles que la tokenisation, au domaine étudié permet une amélioration significative des performances. D'autres chercheurs ont exprimé un fort intérêt pour l'extraction de ces informations à partir de récits cliniques (Bramsen et al., 2006; Hripcsak et al., 2009). Ils ont montré que la notion d'événement est différente quand elle est utilisée dans le domaine général que lorsqu'elle l'est dans le domaine médical.

4.7.3 Challenges des textes cliniques dans l'expression de la temporalité

Pour nos recherches, l'approche mise en œuvre a été choisie en tenant compte de la particularité des comptes rendus. D'après nos échanges avec les médecins, il s'avère que le personnel soignant ne s'intéresse pas à tous les événements qui apparaissent dans le texte mais uniquement à ceux qui sont médicalement pertinents. De plus, lors de la rédaction des comptes rendus, les médecins ont tendance à regrouper les informations à caractères temporel dans le même bloc d'informations comme c'est le cas pour les antécédents qui sont souvent retranscrits dans la première partie du document. Ces événements sont fréquemment liés à une expression temporelle (une date). Des exemples de telles expressions issues des comptes rendus sont illustrés dans les phrases suivantes.

- **01/1999** : sein G T2N1b : tumorectomie + CA : CCI, 20 mm, grade I, emboles, RH+, 12N+/24
- **Dec 2011** : progression pleurale, CA 153 a 178 : pose d'un PAC pleural ; TAXOL hebdo interrompu en **avril 2012** pour toxicité
- **18/03/2016** : syndrome pied/main de grade III/IV, suspension du XELODA
- **03/08** : Xeloda 1800 mg/j 2 sem/3.
- **05.2017** : Hospitalisation en **mai** dernier pour une hematurie macroscopique
- Stabilite de l'imagerie par rapport au **29 mai 2019**
- **Avril 2014** : progression tumorale pulmonaire : TAXOTERE.
- **08/2014** : progression osseuse : irradiation hemibassin gauche ; **10/2014** : GEMZAR (arret en **mai 2015** pour toxicite hematologique).
- **Juillet 11** : decouverte d'un Ca 15-3 a 2N.
- Patient né en **1951**
- La patiente rentre l'**annee prochaine** dans le cadre du depistage systematique organise au niveau mammographie.

Après analyse de plusieurs comptes rendus, aucune forme standard pour l'expression des dates utilisées par les médecins n'a été identifiée. En effet, chaque médecin a la totale liberté d'utiliser les formulations qu'il souhaite lors de la rédaction d'un compte rendu. De plus, les médecins ont tendance à employer des abréviations dans leurs rapports, par exemple «Dec» au lieu de «Décembre». Les séparateurs utilisés entre les jours, les mois et les années sont différents («.», «/», « », etc.). Les jours ne sont pas toujours précisés. Dans l'expression «05.2017», le 05 fait référence au mois de mai. Cependant, certaines formes de ces expressions telles que «03/08» pourrait être difficiles à désambiguïser.

4.7.4 Approche proposée

L'approche que nous proposons consiste en une solution hybride. La première étape est basée sur des règles pour la détection des expressions temporelles. La seconde étape

s'appuie sur un modèle de classification automatique de phrases pour la détection des phrases dites *antécédent* (cf. chapitre 5). La première étape est basée sur des expressions régulières dans l'ordre suivant :

1. Jour-mois-année, soit avec les mois écrits avec des chiffres ou des lettres.
2. Jour-mois, soit avec les mois écrits avec des chiffres ou des lettres.
3. Mois-année, soit avec les mois écrits avec des chiffres ou des lettres.
4. Année seulement, mais dans ce cas, les années sont associées à un contexte. Par exemple «depuis 2008».

Si la date détectée est au début de la phrase avec un format approprié (exemple 1), alors elle est utilisée pour dater tous les concepts non datés de la phrase. Les expressions régulières utilisées sont dans l'annexe C.

4.8 Désambiguïsation sémantique

4.8.1 Désambiguïsation des protocoles

4.8.1.1 Définition du problème

Un protocole de chimiothérapie est un parcours de soins pour un patient. Il définit le nombre de cures et le type de chimiothérapie qui sera utilisé. Un protocole porte un nom court comme FOLFOX ou FEC. Ce nom est souvent un acronyme des différentes molécules utilisées durant le traitement. Lorsqu'on détient la liste de ces protocoles leur identification devient une tâche facile. Cependant, certains de ces protocoles sont ambigus et leur identification induit en erreur les algorithmes à cause des faux positifs qui sont détectés aussi. Les neuf protocoles ambigus sont les suivants : AC, ACE, BEP, CAP, CMF, EP, FAC, FUN et VICE (cf. Annexe D). L'ambiguïté existe avec des termes généraux non spécifiques au projet pour BEP, CAP, FAC, FUN et VICE. Les termes restants sont ambigus avec des termes propres au domaine médical. AC et ACE peuvent désigner des biomarqueurs. EP est une abréviation signifiant « Embolie Pulmonaire » ou « Épanchement Pleurale ». CMF peut correspondre à une mesure ou un prélèvement. Les protocoles les plus importants à désambiguïser sont AC et ACE. Ce sont ceux qui ont le plus souvent un sens différent de celui des protocoles.

Deux approches sont possibles pour désambiguïser : soit mettre en place des règles, soit utiliser un modèle de reconnaissance d'entités nommées. Les deux approches ont des avantages exclusifs. Un modèle pourra découvrir de nouveaux protocoles, tandis que les règles produiront un résultat prévisible, et seront facilement mises à jour. Cependant des règles peuvent s'ajouter et ont tendance à devenir de plus en plus complexes résultant en un ensemble difficile à maintenir.

4.8.1.2 Challenges et analyse de la solution

Une phase d'étude a d'abord été réalisée afin de bien choisir l'approche la plus appropriée. Différents comptes rendus contenant au moins un des neuf termes ont été étudiés. Nous avons alors constaté qu'un protocole ambigu peut apparaître dans tous types de documents. Certains de ces derniers ont été mal océrisés donnant lieu ainsi à

des phrases incompréhensibles et/ou avec une mauvaise tokenisation telle que « FAC TU RATION » à la place du mot « FACTURATION ». Le protocole ACE et le biomarqueur ACE peuvent être employés dans la même phrase. Les protocoles sont parfois mal écrits, on peut lire « l ?ACE », « dAC » ou encore « CA » au lieu de « AC ». Par ailleurs, les médecins parlent indifféremment d'une molécule de chimiothérapie et d'un protocole.

Ces différentes raisons laissent à penser qu'un modèle de reconnaissance d'entités nommées serait plus efficace qu'un ensemble de règles. Certains points communs ont également ressurgi laissant entrevoir des perspectives de réussite pour un modèle. Les protocoles sont majoritairement écrits en majuscules. Les phrases sont courtes et contiennent généralement une date et/ou un chiffre. Il est fréquent que plusieurs protocoles soient évoqués dans la même phrase. Certains types de phrases reviennent fréquemment.

- 1ère ligne métastatique par AC du 24.02.2012 au 23.7.2012
- => mise en œuvre d'une chimiothérapie par 3 cycles d'anthra-cycline type AC et 3 cycles de TAXOTERE du 19.10.2005 au 09/03/2006.
- "Après 5 mois d'imprégnation hormonale le bilan montre une aggravation biologique avec augmentation de l'ACE et du C15.3 à 126 et à 221.
- BEP debute le 18/08/2009.
- Examen clinique et bilan protocolaire J8 C1 dans le protocole ACE.
- CA 15 3 et ACE normaux
- J15 C2 BEP dans le cadre de la prise en charge d ?une tumeur testiculaire germinale non séminomateuse .

4.8.1.3 Approche proposée

Dans le but de former une modèle NER, il faut disposer d'un jeu d'entraînement. 30 000 phrases contenant au moins un des neuf protocoles ont été extraites de l'ensemble du corpus pour constituer les jeux de données nécessaires.

Environ 300 phrases ont été annotées avant d'entraîner le modèle. Lors de cette tâche d'annotation, tous les protocoles présents ont été annotés, et pas seulement ceux qui sont ambigus. Cela permet au modèle de comprendre ce qu'est réellement un protocole. L'emphase est cependant sur les protocoles ambigus car les phrases extraites contiennent au moins un d'eux. Le but de cette première étape est seulement de valider l'approche et non d'aboutir à un modèle définitif. Le jeu de données est divisé en deux parties égales, une pour l'entraînement, l'autre pour le test. Cette première approche nous a permis d'obtenir une exactitude de $71,8\%$ avec une précision de $19,73\%$ un rappel de $22,48\%$ et un F-score de $21,01\%$. Le modèle présente des scores encourageants. En renforçant les annotations, le modèle pourra s'améliorer et obtenir de meilleures performances. Cependant, en testant certaines phrases parmi les 30 000, nous avons constaté un nombre important de faux positifs. Le modèle identifie des protocoles qui n'existent pas.

- NOM DE NAISSANCE : XXXXXX NOM D'EPOUSE : XXXXXX

« NOM DE » est considéré comme un protocole avec une confiance de 99%.

- `Dysgénésie importante (a l'impression que tout ce qu'elle mange est sucre).`

« Dysgénésie importante », « a » et «) » sont considérés comme des protocoles avec respectivement des scores de confiance de 86,77%, 98,54% et 83,11%.

Pour pallier ce problème, une annotation complète a été réalisée. Une annotation est dite complète si toutes les entités de la phrase sont annotées. Ainsi, tous les mots présents dans le jeu de données et non annotés sont considérés comme des NON-PROTOCOLES ajoutant donc beaucoup de données d'entraînement pour le modèle. Une seconde phase d'annotation a été réalisée en plusieurs itérations pour porter le nombre total d'annotations à environ 1000 phrases. Nous avons entraîné plusieurs modèles en utilisant plus de 500 phrases dont 20% utilisées pour le test. On obtient les résultats présentés dans le Tableau 4.1.

Suite aux 1000 annotations et après vérifications dans les 30 000 phrases issues du corpus, aucune ne comprenait le protocole VICE ou FUN, et seulement quelques-unes le protocole EP. Le protocole FUN était correctement détecté par le modèle contrairement aux deux autres. Pour les protocoles VICE et EP une dizaine de phrases ont été créées et rajoutées manuellement pour que le modèle puisse détecter ces protocoles. Après la réalisation d'autres tests sur le modèle, il est apparu que le protocole ACE n'était pas bien détecté. Le biomarqueur ACE apparaît dans un contexte similaire au protocole ACE et souvent les biomarqueurs étaient détectés en tant que protocoles. Pour améliorer la détection du protocole ACE les mêmes phrases ont été ré-annotées avec l'entité BIOMARQUEUR. Cela permet au modèle d'utiliser l'entité BIOMARQUEUR afin de mieux différencier les protocoles. Les résultats obtenus sont présentés dans le tableau 4.1.

Enfin, pour valider les modèles, nous les avons testés sur un gold standard qui doit être annoté par un expert et être intransigeant, contrairement à un jeu d'entraînement qui peut être constitué pour simplifier la tâche d'apprentissage au modèle. Le gold standard ne doit pas être réalisé par la personne ayant constitué l'ensemble d'entraînement, sinon l'annotation risque d'être biaisée. Le gold standard influe l'entraînement du modèle puisqu'il est utilisé comme un jeu d'évaluation. Les résultats obtenus sont présentés dans le tableau 4.1.

4.8.1.4 Résultats

Dans le tableau 4.1 sont présentés tous les résultats obtenus lors des différentes expérimentations effectuées dans le cadre de la désambiguisation des protocoles.

Table 4.1 – Résultats de la désambiguïsation des protocoles.

#Phrases	Hyperparamètres	Ensemble de test	Gold standard
300	0.2 dropout 10 itérations	Exactitude : 84,30 % Précision : 96,08 % Rappel : 24,03 % F-score : 38,51 %	Exactitude : 62,57 % Précision : 77,72 % Rappel : 18,83 % F-score : 30,38 %

500	0.2 dropout 20 itérations	Exactitude : 88,20 % Précision : 94,25 % Rappel : 89,13 % F-score : 91,62 %	Exactitude : 85,47 % Précision : 82,23 % Rappel : 89,30 % F-score : 85,62 %
1000	0.2 dropout 20 itérations	Exactitude : 91,90 % Précision : 93,94 % Rappel : 84,94 % F-score : 91,63 %	Exactitude : 86,20 % Précision : 91,47 % Rappel : 89,77 % F-score : 90,61 %

Pour chacune des expérimentations, plusieurs modèles ont été entraînés en faisant varier les paramètres d'entrée ou les hyper-paramètres à chaque exécution (recherche gloutonne). Dans ce tableau, on ne présente que les performances des meilleurs modèles.

4.8.2 Désambiguïsation des métastases

4.8.2.1 Définition du problème

Les métastases sont reconnues dans un document à l'aide des termes comme « métastases » ou « évolution métastatiques ». Cependant un terme détecté peut ne pas correspondre à une métastase mais à une récidive. Une récidive est à l'instar des métastases une prolifération de tumeur maligne. Une récidive a lieu lorsque les tumeurs sont localisées dans des ganglions ou dans une zone proche du siège de la tumeur primaire. Une métastase correspond à la présence de tumeurs malignes dans des organes éloignés du siège de la tumeur primaire. Ainsi tous les termes indiquant une métastase sont ambigus. Il existe plusieurs termes pour décrire une métastase ou une récidive qui sont employés lorsqu'il y a seulement une suspicion de métastases ou dans un tout autre contexte. Il est important de les détecter car les médecins utilisent rarement les termes simples tels que « métastase » dans les comptes rendus. Les termes ambigus comme « lésion » ou « nodule » sont plus fréquemment utilisés. En effet les médecins analysent des résultats d'imagerie et notent ce qu'ils aperçoivent ; ils ne sont jamais sûrs de voir une métastase mais sont sûrs d'observer une lésion ou un nodule. Enfin certains comptes rendus peuvent être remis au patient, dans ce cas les termes indiquant clairement que le patient présente des métastases peuvent avoir un impact fort, ils préfèrent donc éviter de les employer.

Comme pour la désambiguïsation des protocoles, il est possible de détecter les métastases avec un modèle de NER (Nadeau and Sekine, 2007) ou avec des règles ou un mélange des deux. Cependant, cette fois-ci, il n'y a pas de liste établie de termes.

4.8.2.2 Challenges et analyse de la solution

Une étude des comptes rendus a d'abord été réalisée afin de mieux comprendre la manière dont les médecins évoquent la présence ou non de métastase. Cela a également permis d'extraire une liste de termes ambigus pouvant désigner des métastases, cette liste est illustrée dans la Figure 4.1. Parmi cette liste certains mots sont plus utilisés que d'autres. On peut voir que le terme « lésion » est très employé par rapport aux

autres. Il serait anecdotique de traiter des termes tels que « progression lésionnelle » et « caractère malin » vu leur présence dans le corpus. Ces différents de termes utilisée peuvent être propres à chaque centre mais il est peu probable que les différences soient radicales, Ceux sont les termes les plus simples qui sont les plus utilisés.

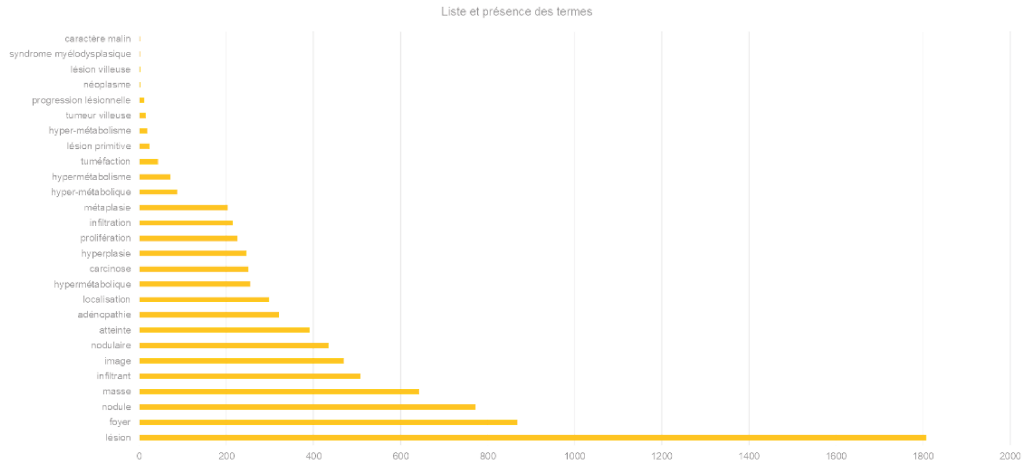


FIGURE 4.1 – Ensemble des termes ambigus pour désigner une métastase et leur présence dans le corpus.

Au cours de cette étude, il a été constaté que les termes ambigus apparaissent souvent dans les comptes rendus qui ont vocation de confirmer ou de réfuter la présence de métastases. Dans ce type de documents (exemple dans 4.2), un terme ambigu n'apparaît généralement pas seul. S'il y a des métastases elles sont probablement présentes à plusieurs endroits.

Contenu

Prescripteur : Dr XXXXX XXXXX (Centre Léon Bérard)
Date de prélèvement : 16/11/2010
Nature de prélèvement : Ponction osseuse crête iliaque gauche - sous scanner -

Renseignements cliniques : lésion lytique crête ilaque.
Antécédent de carcinome mammaire sein droit pT2N0 (2003) traité par chirurgie-chimiothérapie puis hormonothérapie.
Actuellement, la patiente présente de violentes douleurs de la crête iliaque gauche de type inflammatoire depuis environ 1 mois la reveillant la nuit.
Scintigraphie osseuse : lésion suspecte de la crête iliaque avec deux foyers ponctuels sacro iliaque gauche et ischion droit.

EXAMEN MICROSCOPIQUE :

Le matériel de cytoponction a été examiné sur 5 étalements puis sur une cytospine colorés au May Grunwald Giemsa.
La matériel est cellulaire, il est constitué par de nombreux amas carcinomateux tridimensionnels se détachant sur un fond hémorragique. Les cellules sont cohésives au cytoplasme peu abondant, à limites peu distinctes. Les noyaux sont ovalaires renfermant un nucléole central bien visible.

Examen immuno-cytochimique :
Il a été réalisé sur des cytospines.
ER : rares noyaux faiblement marqués.
PR : exceptionnels noyaux, faiblement marqués.

Cytoponction crête iliaque gauche (sous scanner) : **matériel évocateur d'une métastase d'un adénocarcinome, montrant en immuno-cytochimie de rares noyaux tumoraux exprimant faiblement la PR et d'exceptionnels noyaux tumoraux avec un signal faible aux ER.**
Ces aspects réalisés sont compatible avec l'origine mammaire proposée.

Codification ADICAP : CCLO0M40SG

FIGURE 4.2 – Compte rendu confirmant la présence de métastases chez un patient.

Le contexte de la phrase ne suffit pas pour dire si celle-ci parle d'une métastase ou pas. C'est en lisant le document en entier que l'on peut comprendre la signification des phrases contenant un terme ambigu.

Sur 67 phrases issues de 60 documents différents, il était possible de désambiguïser 20 termes en s'appuyant sur le contexte de la phrase, 40 nécessitaient le contexte du document, 2 ont pu être désambiguïsés en consultant d'autres comptes rendus appartenant au même patient, et 5 termes n'ont pas pu être désambiguïsés. Suite à ces observations la mise en œuvre d'un modèle NER ne semblait pas pertinente.

4.8.2.3 Approche proposée

L'approche de désambiguïsation mise en place est alors fondée sur la détection des termes ambigus à l'aide de règles à la fin du traitement du compte rendu. Ces règles sont strictes afin de réduire au maximum le bruit quitte à perdre des informations. Les termes ambigus pouvant être présents partout et surtout avant qu'une métastase soit réellement trouvée.

Chaque terme ambigu possède plusieurs axes d'ambiguïté. Certains sont plus souvent employés pour une signification plutôt qu'une autre. Ainsi les probabilités associées à chaque axe ne sont pas égales.

Trois axes ont été identifiés :

- ORIGINE TUMORALE : la probabilité que le terme désigne une tumeur. Elle sert pour des termes comme « image » ou « foyer » pouvant désigner autre chose par exemple « retour au foyer pour le patient »
- BENIN – MALIN : la probabilité que le terme soit employé pour une tumeur maligne. L'inverse indique la probabilité que le terme soit employé pour une tumeur bénigne.
- PRIMAIRE – SECONDAIRE : la probabilité que le terme soit employé pour une tumeur secondaire (récidive ou métastase). L'inverse indique la probabilité que le terme soit employé pour une tumeur primaire.

Un référentiel « tumeur ambiguë » contenant la liste des termes ambigus identifiés précédemment avec les trois probabilités associées a été créé. Toutes les probabilités ont été fixées à 0,5 sauf pour le terme « lésion». À la détection de chaque terme, la probabilité correspondante lui sera associée. Lors de la phase de désambiguïsation, après le traitement de chaque phrase, les concepts passent par un moteur de règles.

1. La première règle considère les termes ambigus compris dans la même phrase qu'une tumeur primaire. Ainsi la phrase « cancer du sein avec localisation osseuse » remontera une tumeur primaire « cancer du sein » et une métastase « localisation osseuse ».
2. La deuxième règle ne conservera que le terme « lésion » car c'est le terme ayant le plus d'impact (avec une probabilité supérieure à un seuil fixé), et il ne sera remonté que s'il est associé à un concept de localisation ou de traitement. Cette deuxième règle n'est exécutée que si la première n'a pas pu l'être.

Après analyse des dossiers patients en erreur, il est apparu que les dates indiquées dans des fiches tumeurs (documents structurés) correspondent aux dates attendues. Une fiche tumeur est un fichier XML contenant des informations à propos des tumeurs diagnostiquées au patient. Le type de la tumeur, sa localisation, sa morphologie et la date de son diagnostic sont indiqués. Dans le cas d'un cancer, la date de la récidive métastatique peut être indiquée. C'est cette dernière qui nous intéresse. Un dossier patient peut contenir plusieurs fiches tumeurs. Les données contenues dans la fiche tumeur ont un poids plus important, car considérées comme fiables, que les données trouvées dans un compte rendu.

3. La troisième règle consiste à considérer uniquement la date de la fiche tumeur lorsqu'il y en a une pour structurer une métastase. Cependant, cette information n'est pas toujours précisée dans les fiches tumeur.
4. La quatrième règle est lancée seulement quand on ne dispose pas d'une date de début de métastase fiable. Pour cela, toutes les dates associées à des concepts de métastases identifiées sur une période d'un an pour un patient sont parcourues. La première date (dans le temps) sera celle retenue pour le début de la métastase. Ensuite un score de complétude est calculé indiquant la fiabilité de la date. Ce score a été établi empiriquement après observations des scores de complétudes de différents patients.

4.8.2.4 Résultats

Le but de la mise en place de cette désambiguïsation est d'améliorer la détection des métastases. Afin d'évaluer l'intérêt et la pertinence de la détection des termes ambigus et d'évaluer correctement l'ensemble des règles, le jeu de tests a été composé d'une quarantaine de patients. Ces règles ont permis de détecter les métastases correctement pour 27 patients, et de rapprocher la date de détection de métastase, sans être juste, pour 7 patients.

Des améliorations sont ainsi observables laissant penser à un apport bénéfique. De plus, lors de l'étude effectuée sur ces patients, il a été remarqué que certaines dates attendues étaient fausses dans le fichier de validation. Ainsi, des corrections ont été proposées et envoyées à des médecins du centre de cancérologie de Lyon pour confirmation. Une fois cette dernière obtenue, les nouvelles dates ont été mise à jour dans le fichier de validation.

4.8.3 Désambiguïsation des codes SNOMED-CT

SNOMED-CT (*Systematized Nomenclature of Medicine-Clinical Terms*) est un système de terminologie standard. Il s'agit d'une terminologie clinique hiérarchique des soins de santé contenant des termes médicaux et leurs relations. SNOMED-CT contient des signes cliniques (symptômes), des troubles (diagnostics), des procédures, des structures corporelles, des organismes, etc. Cette terminologie est souvent utilisée par les médecins sous forme de codes dans les comptes rendus. Ces codes sont souvent représentés sur cinq chiffres tout comme les codes postaux des adresses en France.

Contrairement aux codes SNOMED-CT, que l'on retrouve souvent au milieu ou en fin des comptes rendus, les codes postaux sont généralement utilisés au début des rapports dans la partie dédiée aux renseignements personnels du patient quand elle existe. Savoir distinguer alors entre une phrase à caractère administratif et une phrase contenant des informations médicales pourrait aider à désambiguïser le sens de ces codes quand ils sont détectés. Dans le chapitre 5, nous allons aborder ce processus de distinction entre les phrases selon leur contenu sous forme de classification automatique des phrases.

Par ailleurs, compte tenu des difficultés liées à l'utilisation de certaines normes et la détection des codes morphologiques de façon correcte, le standard SNOMED-CT a été décomposé en plusieurs lexiques : noms de tumeurs, adjectifs de tumeurs, adjectifs généraux, et comportements tumoraux. Ce choix a pour but de développer autant de descriptions de modèles morphologiques qu'il y en a dans les différents types des comptes rendus.

4.9 Détection de la négation et de l'incertitude

4.9.1 Définition du problème

L'identification des phrases négatives et incertaines dans les textes est une thématique de recherche qui a fait l'objet de plusieurs études appliquées à des domaines variés ces dernières années (Chapman et al., 2001; Mutalik et al., 2001; Packard et al., 2014; Fancellu et al., 2016). En effet, dans le traitement automatique de langage naturel, les données textuelles sont souvent non structurées. La détection de telles phrases joue alors un rôle prépondérant dans les tâches de recherche et d'extraction d'informations.

Dans le domaine médical, les médecins utilisent souvent la négation pour exclure un diagnostic ou un traitement. Ils utilisent aussi des formulations hypothétiques sous forme d'incertitude afin de souligner la prudence avec laquelle ils souhaitent s'exprimer sur l'existence ou l'absence d'un fait. La détection de phrases négatives ou incertaines permet alors d'affirmer/infirmer les concepts identifiés dans la même phrase. Cela joue un rôle important et impacte significativement les tâches de fouille de texte, notamment lors de la sélection de patients pour la création de cohortes de patients homogènes ou lors de reconstitution du parcours palliatif des patients à partir de leurs comptes rendus.

4.9.2 Challenges et analyse de la solution

Les expressions utilisées par les médecins pour poser un diagnostic sont parfois difficiles à comprendre. Lorsqu'ils raisonnent sur le diagnostic d'un patient, ils essaient d'exclure les symptômes que le patient ne présente pas. Pour cela, ils utilisent ce qu'on appelle des déclencheurs de négations qui sont répartis sur trois types : pré-négation, post-négation et pseudo-négation. Il existe dans la littérature plusieurs travaux qui proposent des listes plus ou moins exhaustives de ces déclencheurs. Le Tableau 4.2 regroupe quelques exemples de déclencheurs de négation.

TABLE 4.2 – Liste de déclencheurs de négation en langue française.

Adjectifs	absent, aucun, négatif, nul,...
Adverbes	jamais, ne pas, ne [...] pas, ne [...] rien, non, pas, ...
Affixes	a-, dis-, im-, in-, non-,...
Conjonctions	ni, ni [...] ni [...], sans que, ...
Locutions	à l'exception, au lieu de, ...
Noms	absence de, ...
Prépositions	excepté, sans, sauf, ...
Pronoms	personne, rien,...
Verbes	annuler, éliminer, exclure, nier,...

Bien que la détection des déclencheurs de négation constitue des difficultés en raison de l'ambiguïté et du contexte, il est bien plus difficile d'identifier leurs portées avec précision. L'identification de cette portée revient à répondre à la question : qu'est-ce qui est réellement nié ? Une phrase négative pourrait être constituée d'un nombre important de mots, alors que la négation porte seulement sur une sous-partie de la phrase. La portée d'une négation peut se trouver à différentes endroits dans la phrase : avant le déclencheur, après le déclencheur ou avant et après le déclencheur. En plus, compte tenu de la diversité et de la richesse que peut avoir une langue, un seul déclencheur peut avoir à lui seul plusieurs portées dans la même phrase. Une phrase pourrait très bien contenir plusieurs marqueurs ou déclencheurs de négation. Une autre difficulté à gérer serait alors d'associer chaque portée à son propre déclencheur. Pour compliquer davantage les choses, il existe par ailleurs des exceptions à prendre en compte telles que les doubles négations. Quelques exemples extraits des comptes rendus sont illustrés dans les phrases suivantes.

- PS : **0**
- Stroma fibreux : **Absent**
- Transfusion plaquettaire :**Non**
- **Pas** d'anomalie à l'examen gunécologique.
- **Absence** de lésion tumorale.
- L'étude en fenêtre osseuse **ne** met **pas** en évidence de lésion focale suspecte.
- Il **n**'existe **aucune** anomalie mammographique.
- L'examen clinique **ne** montre **rien** de péjoratif
- pour le moment compte tenu de la situation un retour en centre est **exclu**.
- Pièce mesurant 1 x 0,5 cm ayant fait lobjet d'un examen extemporané (Dr CHETAILLE) qui est répondu **négatif**
- On demande que les coupes thoraciques soient fait en protocole angioscanner afin d'**éliminer** l'embolie pulmonaire.
- Patiente de 63 ans, menaupausee **sans** THS, vue ce jour dans le cadre de la consultation BILBAO pour bilan d'extension d'une neoplasie mammaire droite.

Cependant, la présence d'un déclencheur d'une négation dans une phrase n'implique pas forcément l'annulation des concepts couverts par sa portée. Certaines phrases négatives peuvent véhiculer des informations très importantes notamment dans le cadre de notre recherche. Parmi les objectifs initiaux de notre thèse l'identification des pa-

tients résistants au traitement anticancéreux. Or, ces résistances sont modélisées sous formes de réponses au traitement. La phrase : «Pas d'évolutivité» est traduite par une *stabilité* qu'est l'une des quatre niveaux de réponse en oncologie (cf. chapitre 6).

- cancer primitif oui ( ) **non** ( )
- Le PSA était à 0.01 en octobre 2017, mais le ZOLADEX **n**'a été refait **qu**'en février 2018.

De la même manière, les médecins utilisent souvent des phrases spéculatives ou incertaines. À la lecture seule d'une phrase incertaine, on est pas capable de répondre par oui ou non à tous les concepts ou sujets évoqués dans celle-ci. Contrairement aux négations, il est très difficile de concevoir une liste exhaustive regroupant l'ensemble des déclencheurs ou marqueurs de l'incertitude. En effet, en langue française, il est possible d'employer le temps conditionnel afin d'exprimer l'éventualité d'un événement. Cela rend les approches basées sur de simples règles et des listes de déclencheurs obsolètes. Plusieurs exemples de phrases incertaines issues des comptes rendus sont ci-dessous.

- Toujours est-il qu'il **semble** exister sous POMALIDOMIDE DEXAMETHAZONE des arguments évoquant **potentiellement** une infection des voies respiratoires
- Un suivi gériatrique en période post-opératoire qui **peut être** assuré par l'équipe mobile de gériatrie de l'hôpital Saint Luc Saint Joseph (Dr Arnoux)
- Donner une contre indication absolue et definitive aux Antracyclines me **parait** dommageable **si** c'est le traitement le plus adapte aux ennuis de sante actuels de Mme XXXXXX
- Il existe une indication de **recherche** de mutation constitution-nelle des gènes BRCA1/BRCA2 chez Mme XXXXXX, discutée et validée sur dossier avec le Docteur NOGUES
- Grand-mère décédée à 47 ans, tumeur pulmonaire (cancer primitif **?**)
- **Hypothèse** la plus **probable**, l'existence de facteurs de **susceptibilité** de nature génétique unique **ou** multiple en association
- J'informe Mme XXXXXX qu'**en cas** de prédisposition génétique avérée, un test **pourra être proposé** à ses apparentés
- un **éventuel** foyer de myosite circonscrite vu l'antécédent
- Devant ces malaises atypiques, on peut se poser la **question** d'une toxicité de l'Asparaginase
- Grand-mère décédée à 27 ans d'une **possible** pneumonie
- Je reste néanmoins très circonspecte sur la qualité de cette rémission et sur la nécessité d'une grande prudence et d'une surveillance étroite en raison du **risque** de rechute
- L'**alternative possible** de mastectomie prophylactique dans ce contexte a été abordée
- La patiente est **potentiellement** incluable dans l'essai PAOLA1
- Les options thérapeutiques évoquées **dans le cas** d'une absence de progression tumorale étaient **soit** une cystoprostatectomie-curage, **soit** une cystoscopie-résection suivie d'une radio-chimiothérapie ; ces **hypothèses** thérapeutiques ont été évoquées avec M. XXXXXX ce jour
- Les recoupes en série effectuées ne mettent pas en évidence d'éléments **suspects** de malignité
- Ponction : adenocarcinome, **doute** sur metaplasie
- Formation pelvienne droite entre lovaire droit et l'utérus tubulée d'allure digestive contenant une image dense **pouvant** faire évoquer un stercolite

À l'instar des phrases négatives, les phrases incertaines sont aussi concernées par la portée des déclencheurs d'incertitude.

4.9.3 Approche proposée

La totalité des approches utilisées pour la détection des phrases négatives et/ou incertaines ainsi que leur portées respectives que nous avons pu consultées dans la littérature sont réparties selon trois catégories. Des approches basées sur des règles à base d'expressions régulières et des dictionnaires, des approches basées sur des techniques d'apprentissage automatique, et des approches hybrides.

Afin de répondre à cette tâche, une méthode originale composée de deux étapes a été proposée. La première étape consiste en une tâche de classification automatique des phrases à l'aide d'un modèle d'apprentissage automatique pour la détection des phrases négatives/incertaines. Cette première étape constitue une couche de raffinement afin de ne cibler que les phrases concernées par la négation ou l'incertitude, et ainsi éviter de chercher des portées dans toutes les phrases du corpus. Même s'il est relativement facile d'utiliser les expressions identifiées dans le Tableau 4.2 pour identifier les phrases négatives, il existe de nombreuses formulations de négation difficiles voire impossible à détecter via des règles. Ces exceptions sont parfois liées aux particularités des textes cliniques.

De plus, comme déjà évoqué, il n'existe aucune liste exhaustive pour les déclencheurs de l'incertitude. L'établissement d'une telle liste est une tâche laborieuse et fastidieuse. La deuxième étape est dédiée à la détection des portées des négations ou des incertitudes. Compte tenu de la qualité des données à traiter, de la richesse de la langue française, et de la complexité de cette tâche, la deuxième étape a été modélisée comme une tâche de reconnaissance d'entités nommées. Un modèle NER a été entraîné sur plusieurs milliers de phrases pour reconnaître la portée des négations et des incertitudes. Cette approche fournit des résultats très satisfaisants. En plus des bonnes performances, l'un des avantages de cette méthode est que le modèle NER (Nadeau and Sekine, 2007) permet non seulement de détecter les portées, mais aussi de les rattacher aux déclencheurs appropriés. Les détails techniques, les expérimentations et les résultats de cette approche seront abordés dans le chapitre 5 suivant.

4.10 Conclusion

Dans ce chapitre, nous avons mis l'accent sur les opérations de pré-traitements effectuées sur les comptes rendus médicaux afin de rendre les connaissances qui y sont contenues accessibles. Nous avons démontré l'utilité de ces tâches dans le processus de fouille de texte et du traitement automatique de langage naturel. Plusieurs techniques ont été présentées : de la collecte et la constitution du corpus à la détection de phrases négatives et incertaines en passant par la tokenisation, la détection de limites de phrases, la normalisation (lemmatisation, correction orthographique, etc.), la détection des dates et enfin la désambiguïsation sémantique. Les approches proposées ont été choisies en tenant compte à la fois de la complexité de la tâche, de la particularité des comptes rendus traités, mais aussi en gardant à l'esprit la nature des traitements

qu'on souhaite effectuer et les informations qu'on aimerait extraire pour répondre à la problématique générale à savoir l'identification et la caractérisation des patients résistants aux traitements d'oncologie.

Par ailleurs, les référentiels intégrés (cf. Annexe A) permettent d'identifier plusieurs entités dans les comptes rendus telles que : les médicaments de chimiothérapie, les molécules utilisées, les protoboles, les biomarqueurs, les localisations et les côtés d'une tumeur (par exemple : «cancer du [1]**sein** [2]**gauche**»), les comportements de tumeur, les gènes, les actes, les diagnostics, etc. Ensuite, à partir de ces concepts, de nouvelles entités composites sont créées telles que : les cancers, les évolutions tumorales, les lignes et les cycles de traitement, etc. Ces entités seront discutées plus en détails dans le chapitre 6.

1. localisation : sein
2. côté : gauche

CHAPITRE 5

CLASSIFICATION

Sommaire

5.1 Introduction

La classification de textes est un processus qui consiste à attribuer à un texte une ou plusieurs catégories pertinentes à partir d'un ensemble prédéfini en fonction de son contenu. C'est l'une des tâches fondamentales du traitement automatique de langage naturel et de l'analyse de textes. Les classificateurs de textes sont utilisés afin d'organiser, de structurer et de catégoriser des corpus de données souvent très larges. La classification de textes ou des phrases a été utilisée dans des applications diverses et variées telles que l'analyse des sentiments, la catégorisation des sujets, la détection de spams ou la détection d'intention. Les textes non structurés constituent l'écrasante majorité des données textuelles, et représentent une source d'informations extrêmement riche. Cependant, l'extraction de ces informations peut être difficile et parfois chronophage en raison de la nature non structurée de ces textes.

Dans la littérature, la classification automatique de texte implique trois types de systèmes différents : (i) Les approches basées sur des règles ; (ii) Les approches basées sur l'apprentissage automatique ; et (iii) Les approches hybrides.

Dans ce chapitre, nous allons présenter plusieurs modèles d'apprentissage automatique pour la classification des phrases issues de comptes rendus médicaux, en se basant sur des algorithmes d'apprentissage appartenant à des familles différentes. Le chapitre 4 précédent a fait l'objet d'études des différentes tâches de pré-traitement appliquées aux comptes rendus utilisés dans notre thèse. Ces pré-traitements permettent de rendre ces documents facilement accessibles et prêts à être utilisés par des processus unifiés. La classification de phrases est l'un de ces processus, et est constituée de quatre étapes complémentaires à savoir : la représentation et l'extraction de caractéristiques, la réduction de dimensions, la sélection et la formation de classificateurs, et enfin l'évaluation de ses performances.

Nous allons étudier dans ce qui suit les différentes étapes de classification de textes appliquées aux comptes rendus médicaux. Plusieurs algorithmes d'apprentissage automatique seront utilisés pour la réalisation des expérimentations. L'ensemble des expérimentations effectuées seront présentées et les résultats obtenus seront analysés.

5.2 Extraction de concepts médicaux

Dans le chapitre 3, nous avons expliqué le fonctionnement de la chaîne de traitement des comptes rendus. Plusieurs référentiels sont utilisés afin d'extraire et d'enrichir les concepts médicaux à partir des textes. Ces concepts vont être ensuite exploités pour structurer les documents dans un premier temps, et les patients dans un second temps. Le processus d'extraction de concepts est appliqué phrase par phrase selon les catégories qui lui sont attribuées. Conformément aux référentiels utilisés, chaque concept est constitué d'un ou plusieurs *tokens*. Les principaux concepts extraits et leur description sont présentés dans le Tableau 5.1 ci-dessous.

TABLE 5.1 – Liste de quelques concepts extraits dans la chaîne de traitements et leur description.

Concept	Description
Médicament	Les médicaments utilisés lors des traitements de chimiothérapie. Exemples : «aclasta» ou «17-bêta-œstradiol».
Molécule	Les molécules utilisées dans les médicaments de chimiothérapie. Exemples : «bléomycine», «kyprolis».
Protocole	La pratique d'un acte médical ou paramédical, selon une bibliographie, une expérience clinique partagée, ou encore des recommandations d'un consensus de professionnels. Exemples : «FOLFIRINOX», «LV5FU2».
Biomarqueur	Une caractéristique biologique utilisée pour le dépistage médical, le diagnostic (caractérisation d'une maladie chez un individu), la réponse à un traitement médical, la rechute après un traitement, la toxicité d'une molécule. Exemples : «HER2+», «CA15.3».
Tumeur primaire	Tumeur principale à partir de laquelle peuvent s'échapper des cellules cancéreuses qui vont former des métastases dans d'autres parties du corps. Exemples : «adénopathie», «lésion primitive».
Adjectif de tumeur	Les adjectifs utilisés pour la description d'un tumeur. Exemples : «carcinoïdes», «adénosquameux».
Localisation	Localisation d'une tumeur, partie du corps humain. Exemples : «pulmonaire», «ganglions».
Comportement	Comportant de la tumeur. Exemples : «infiltrant», «invasif».
Gène	Unité fonctionnelle d'hérédité qui est un segment d'ADN trouvé sur les chromosomes dans le noyau d'une cellule. Exemples : «kiaa1549», «flj27352».
Acte	Les actes chirurgicaux réalisés lors d'une chimiothérapie. Exemple : «Anastomose spongiocaverneuse, par abord direct».
Diagnostic d'une tumeur	Diagnostic établi après la réalisation d'examens cliniques, biologiques et d'imagerie. Exemple : «*** SU11 *** Leucémie monocytaire subaiguë».

La détection de concepts au sein d'une phrase se fait principalement avec des règles simples et des patterns qui permettent de détecter un mot présent dans un ensemble préalablement défini tel qu'un référentiel.

Contrairement aux différentes démonstrations de traitement automatique de lan-

gage naturel réalisées dans les tâches classiques de fouille de textes, nous sommes confrontés ici à des problématiques de volumétrie liées à la taille des lexiques médicaux utilisés, qui sont extrêmement fournis, et à la quantité de données différentes.

5.3 Représentation des données et extraction des caractéristiques

Afin de pouvoir appliquer les algorithmes d'apprentissage automatique sur l'ensemble des données pré-traitées, les données textuelles doivent être d'abord représentées sous forme de vecteurs numériques pour former l'entrée de ces algorithmes. Cette opération s'appelle la représentation de texte, et est l'un des problèmes fondamentaux de la fouille de textes et de la recherche d'informations (IR). Elle vise à représenter numériquement des documents textuels non structurés afin de les rendre mathématiquement calculables. Pour un ensemble donné de documents texte $D = \{d_i, i = 1, 2, ..., n\}$, où chaque d_i représente un document, la représentation des textes consiste à représenter chaque d_i de D comme un point e_i dans un espace numérique E, où la distance/similarité entre chaque paire de points dans l'espace E est bien définie. Ainsi, plusieurs caractéristiques sur ces textes peuvent être extraites. L'extraction de caractéristiques à partir des données textuelles a un rôle prépondérant dans la classification automatique de ces données.

5.3.1 Représentation mathématique d'un corpus de comptes rendus médicaux

Pour éviter les pertes d'informations sémantiques causées par l'utilisation des modèles de sac de mots pour la transformation des textes des comptes rendus en vecteurs numériques, nos travaux ont été axés sur l'utilisation des plongements de mots (*word embedding*) (Levy and Goldberg, 2014) pour la représentation des textes. En effet, les modèles de sac de mots ne tiennent pas comptes des relations sémantiques qui peuvent exister entre les mots. Cela constitue un réel problème quant à la compréhension du contenu des comptes rendus.

Un modèle de vecteurs de mots a été utilisé, où chaque mot du vocabulaire se voit attribuer un vecteur numérique. Le modèle est considéré comme étant bien formé si les vecteurs proches dans l'espace correspondent à des mots synonymes ou souvent utilisés dans des contextes similaires. L'entraînement d'un tel modèle est une technique d'apprentissage non-supervisé. Les données ne sont donc pas nécessairement annotées. En revanche, pour espérer de bonnes performances, il est nécessaire d'entraîner le modèle sur de grandes quantités de données, et cela pour deux raisons. D'une part, pour gagner en cohérence quant à la valeur des vecteurs résultants, d'autre part, afin d'augmenter la taille du vocabulaire et couvrir ainsi tous les mots (et variantes de mots) utilisés dans les données.

La question du choix de l'algorithme s'est donc posée tout naturellement. Pour répondre à cette question, deux approches de l'état de l'art ont été étudiées et comparées : *Word2Vec* (Mikolov et al., 2013) et *FastText* (Joulin et al., 2016). Bien qu'efficace, la

principale limitation du modèle *Word2Vec* est que l'algorithme ne prend pas en compte la similarité syntaxique des mots. Cet inconvénient est un point important notamment lors du traitement des textes des comptes rendus contenant des mots mal orthographiés, auquel cas les vecteurs résultant devraient être similaires à (ou proches de) ceux des mots sous leur forme correcte. L'algorithme de *FastText* (Joulin et al., 2016) vient justement remédier à ce problème, puisqu'il s'intéresse à la forme syntaxique des mots. Cette seconde approche est donc plus appropriée pour la représentation vectorielle des comptes rendus.

$$similarité = \cos\theta = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \in [-1, 1] \tag{5.1}$$

Où A_i et B_i sont des composants de A et B respectivement, des vecteurs non nuls de n dimensions.

La similitude qui en résulte est comprise dans l'intervalle $[-1, 1]$. -1 signifie que les vecteurs sont résolument opposés, la valeur 0 indique que les vecteurs sont indépendants (orthogonaux) et la valeur 1 signifie que les vecteurs sont similaires (colinéaires de coefficient positif). Les valeurs intermédiaires permettent d'évaluer le degré de similarité. Appliquée à des données textuelles, la similarité cosinus (Tata and Patel, 2007) de deux vecteurs de documents est comprise dans $[0, 1]$. Les vecteurs des documents similaires ont ainsi une valeur de similarité cosinus très faible (proche de 0).

5.3.1.1 Entraînement de modèles de plongements de mots à partir des comptes rendus du corpus et des données de Wikipedia

Dans un première temps, un modèle a été entraîné sur l'intégralité de notre corpus (360.000 comptes rendus). Le but de cette première étape est seulement de valider notre approche et non d'aboutir à un modèle définitif. Une première comparaison avec un modèle entraîné essentiellement sur l'intégralité du corpus Wikipédia français confirme le besoin de construire des vecteurs de mots adaptés au contexte médical. Pour réaliser cette évaluation, 320 termes répertoriés sur 11 catégories ont été sélectionnés. Les distances inter-couples sont calculées pour tous les couples de termes possibles au sein d'une même catégorie, en utilisant la distance cosinus. La moyenne pour chaque catégorie ainsi que la moyenne finale, et le nombre de mots reconnus par le modèle sont donnés pour les deux modèles. Le Tableau 5.2 suivant présente les résultats obtenus.

TABLE 5.2 – Comparaison des scores de distance des deux premiers modèles.

Catégorie	**Modèle Wikipedia**	**Modèle corpus**
Métastase	0.4821	0.4327
Biomarqueur	0.9715	0.901
Ville	0.6619	0.8508
Localisation	0.5093	0.7301
Médecin	0.9281	0.7085
Date	0.5128	0.4676

Unités	0.7792	0.6381
Tumeur primaire	0.86	0.7442
Protocole	0.9513	0.7311
Acte	0.7849	0.4121
Molécule	0.9972	0.7969
Moyenne	**0.8246**	**0.7254**
#Mots identifiés	179	252

D'après les résultats présentés ci-dessus, le modèle entraîné sur les comptes rendus médicaux fournit de meilleurs résultats pour 9/11 catégories comparé au modèle entraîné sur Wikipedia. Bien que ce dernier représente un volume de données beaucoup plus important (plus de 2.000.000 articles qui sont généralement plus grands que les comptes rendus). Cependant, le modèle entraîné sur Wikipedia est meilleur sur des catégories plus génériques, telles que «Ville» et «Localisation» (faisant référence à une partie du coprs humain). Le nombre de mots détectés est bien plus élevé pour le modèle entraîné sur les comptes rendus car la liste des termes a été construite à partir de termes récurrents dans les comptes rendus.

5.3.1.2 Entraînement du modèle de plongements de mots à partir des comptes rendus de l'Institut Curie

Ces observations laissent à penser que l'entraînement d'un modèle sur l'intégralité des comptes rendus d'un centre de lutte contre le cancer pourrait améliorer considérablement les performances des résultats obtenus décrits ci-dessus. Pour ce faire, le choix s'est porté sur l'Institut Curie (IC) puisqu'il recense le plus grand nombre de comptes rendus, plus de 16.000.000 de documents. Ce modèle devrait améliorer les performances sur des termes plus génériques. Les résultats obtenus sont présentés dans le Tableau 5.3 ci-dessous ainsi que les résultats des deux premiers modèles à titre de comparaison.

TABLE 5.3 – Performances du modèle entraîné sur les données de l'Institut Curie sur 320 mots.

Catégorie	**Modèle Wikipedia**	**Modèle corpus**	**Modèle IC**
Métastase	0.4821	0.4327	0.2964
Biomarqueur	0.9715	0.901	0.7587
Ville	0.6619	0.8508	0.5877
Localisation	0.5093	0.7301	0.655
Médecin	0.9281	0.7085	0.805
Date	0.5128	0.4676	0.3899
Unités	0.7792	0.6381	0.6442
Tumeur primaire	0.86	0.7442	0.7106
Protocole	0.9513	0.7311	0.5381
Acte	0.7849	0.4121	0.3186

Molécule	0.9972	0.7969	0.5836
Moyenne	**0.8246**	**0.7254**	**0.5977**
#Mots identifiés	179	252	299

Les résultats obtenus avec le modèle entraîné sur tous les comptes rendus de l'Institut Curie sont bien meilleurs globalement que les résultats précédents et le vocabulaire est plus riche quant au nombre des termes reconnus. Quelques catégories obtiennent toutefois de meilleurs scores dans les modèles précédents, telles que «Localisation», «Médecin» et «Unités». Pour les noms des médecins, ce résultat peut s'expliquer par le fait que les noms utilisés dans l'ensemble de test sont plus présents dans les comptes rendus de notre corpus constitué de 360.000 documents. Pour la catégorie «Unités», les scores obtenus sont légèrement meilleurs dans le modèle entraîné sur notre corpus. Mais de manière générale, le modèle entraîné sur les données de l'Institut Curie est bien meilleur sur des domaines très spécifiques (Biomarqueurs, Protocoles, Chimiothérapies, etc). Cela est dû au grand nombre de données vues lors de la formation des modèles.

Pour évaluer la distribution des distances calculées, des histogrammes ont été construits, pour représenter la concentration des scores obtenus pour différents intervalles, allant de 0 jusqu'à 1 avec un pas de 0.05. La concentration des données vers l'origine (axe des abscisses) est un bon indicateur pour l'évaluation du modèle. Les différentes barres de l'histogramme sont colorées en fonction des différentes catégories auxquelles correspondent les scores (Figure 5.1).

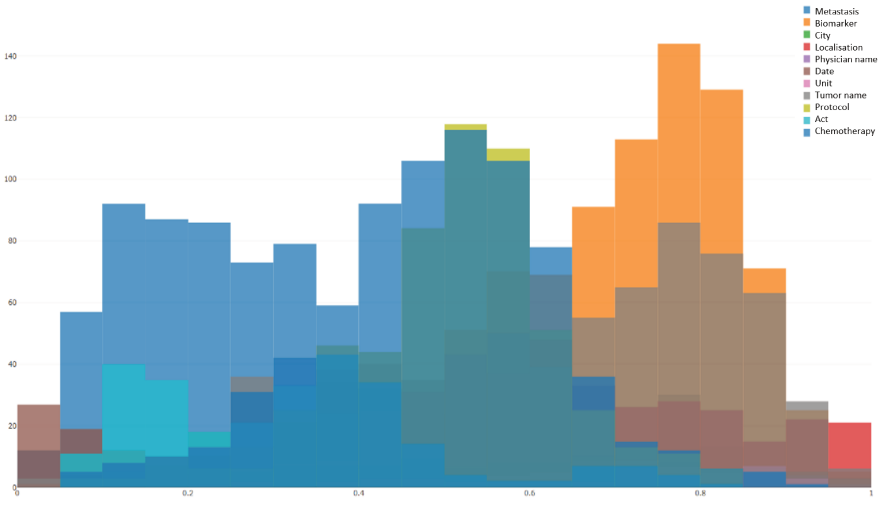


FIGURE 5.1 – Distribution des scores de distance pour le modèle entraîné sur l'IC.

Après l'analyse de la Figure 5.1, on constate que la répartition des scores de distance pour les couples appartenant à la catégorie «Biomarqueur» (en orange) est très concentrée dans l'intervalle (0.65, 0.9), et que peu de couples inter-catégories ont un score supérieur à 0.9, ce qui est bon signe.

Afin de visualiser la cohérence des différents modèles de manière plus spécifique, des représentations 3D ont été réalisées en projetant les vecteurs des 320 termes sur un espace tridimensionnel à l'aide de l'algorithme d'analyse de composantes principales (PCA) (Jolliffe, 1986). Chaque catégorie est représentée par une couleur afin de bien visualiser les clusters correspondant aux différentes catégories. Les représentations correspondant aux modèles entraînés sur les comptes rendus de notre corpus (à gauche) et sur les données de l'Institut Curie (à droite) sont illustrées dans la Figure 5.2.

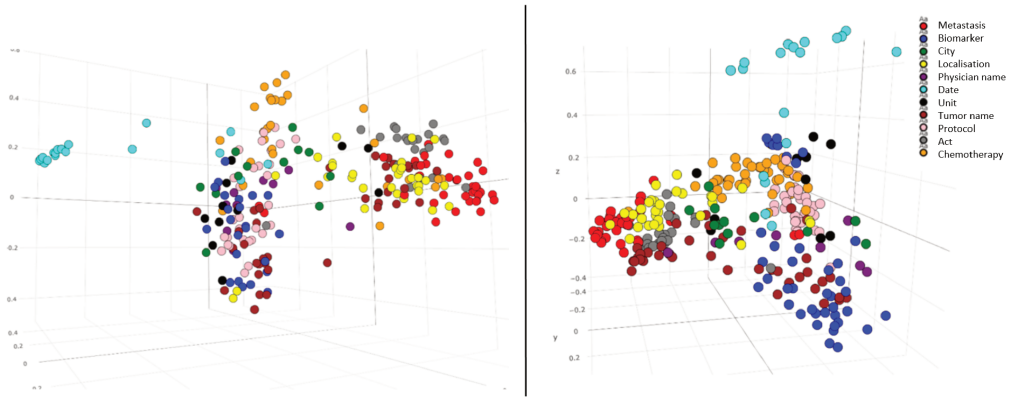


FIGURE 5.2 – Comparaison des représentations 3D de vecteurs de mots avec deux modèles différents (après utilisation de PCA pour la réduction de dimensionnalité).

Bien que l'angle d'observation ne soit pas le même, les deux modèles, ont dans tous les cas, une répartition distincte des vecteurs dans l'espace. Les clusters formés à l'aide du modèle entraîné sur les données de l'Institut Curie (à droite) sont mieux regroupés sur la Figure 5.2. Cependant, ils ne sont évidemment pas formés de façon parfaite, car il faut noter d'une part le grand nombre de catégories proposé entraînant donc un niveau de bruit conséquent, et d'autre part le fait que certains termes appartiennent à plusieurs catégories, ou tout simplement peuvent être utilisés dans des contextes très similaires (par exemple «Biomarqueur» ou «Protocole», concepts pour lesquels de la désambiguïsation sémantique est réalisée). De plus, il est important de relever que le modèle de *FastText* a tendance à effectuer une séparation importante entre les mots en majuscule et en minuscule de par son algorithme leur attribuant des n-grammes totalement distincts. Or, comme cela a été discuté dans le chapitre précédent (4), certains mots sont souvent écrits en majuscule dans certains contextes, tels que les protocoles ou les médicaments. Enfin, les résultats obtenus sont satisfaisants puisqu'ils confirment l'amélioration observée grâce à l'exploitation des données de l'Institut Curie.

5.3.1.3 Intégration de la tokenisation dans le processus de plongements des mots avec *FastText*

Le réel objectif derrière l'entraînement des vecteurs de mots est d'optimiser les performances de la tâche de classification des phrases. Lors du calcul des plongements des mots avec *FastText*, la tokenisation est effectuée en fonction des espaces séparant

les mots. Dans le chapitre des pré-traitements des données textuelles (cf. Chapitre 4), nous avons réalisé que compte tenu des particularités des textes cliniques étudiés, la tokenisation, entre autres tâches de préparation des données, a été adaptée en intégrant l'ensemble des référentiels utilisés lors de l'enrichissement sémantique dans le cadre du projet (cf. Annexe A). Il est donc plus approprié d'entraîner les vecteurs de mots à partir des données traitées avec le même processus de tokenisation. Les résultats obtenus sont présentés dans le Tableau 5.4.

TABLE 5.4 – Performances du modèle entraîné sur les données préalablement tokenisées de l'Institut Curie sur 320 mots.

Catégorie	**Modèle IC**	**Modèle IC + Tokenisation**
Métastase	0.2964	0.4506
Biomarqueur	0.7587	0.8081
Ville	0.5877	0.6985
Localisation	0.655	0.7735
Médecin	0.805	0.8738
Date	0.3899	0.4454
Unités	0.6442	0.7202
Tumeur primaire	0.7106	0.7951
Protocole	0.5381	0.6418
Acte	0.3186	0.5133
Molécule	0.5836	0.7444
Moyenne	**0.5977**	**0.7044**

Bien que les scores obtenus avec notre approche de tokenisation soient toujours inférieurs à ceux obtenus précédemment, l'objectif principal est d'optimiser les performances de la classification de phrases et de la reconnaissance d'entités nommées. En effet, le fait d'utiliser le même processus de tokenisation pour l'entraînement des vecteurs des mots, lors de la détection des limites des phrases, est un point clé pour l'obtention de résultats cohérents. Dans la section 5.4, nous allons étudier l'impact de ce choix sur les performances de la classification des phrases issues des comptes rendus.

5.4 Classification automatique des phrases

5.4.1 Définition du problème

Les comptes rendus représentent une mine d'informations riche en connaissances qui doit être impérativement exploitée afin d'améliorer le processus de soin de santé. Ces documents contiennent une variété d'informations médicales, telles que les décisions cliniques prises par les médecins, les traitements administrés aux patients, les antécédents (qui comprennent, entre autres, les chirurgies antérieures, les hospitalisations et les maladies chroniques des membres de la famille). Mais également des informations de nature administrative telles que le nom de l'institution, les adresses, etc. Pourvoir distinguer, de manière automatique, entre les informations comprises dans chaque phrase permet d'aider les médecins lors de la prise de décision médicale.

La classification automatique des phrases est l'une des techniques de traitement de textes qui permet de tirer parti de ces données. Il s'agit d'une une tâche fondamentale, souvent effectuée en amont des techniques de traitement du langage naturel, et joue un rôle essentiel dans la recherche et l'extraction d'informations. Formellement, étant donné un ensemble $S = \{x_1, ..., x_n\}$ de phrases brutes et k indices prédéfinis de différentes valeurs discrètes, un modèle de classification de phrases attribue l'un ou les indices les plus appropriés pour chaque entrée X_i, en tenant compte à la fois de la forme et de la signification de chaque phrase X_i dans un contexte général. Dans la littérature, de nombreuses approches de classification de phrases ont été proposées, avec des résultats prometteurs. Cependant, pour proposer un modèle efficace, il est important d'adapter la tâche aux particularités liées au domaine d'application afin de mieux appréhender la complexité syntaxique et sémantique des phrases.

5.4.2 Challenges et analyse des données

Une étude des différents types de comptes rendus a été réalisée afin de comprendre la nature des informations qui y sont contenues. Au cours de cette étude il a été remarqué que les comptes rendus diffèrent fortement selon les centres. Il n'y a pas de forme commune. Par exemple un centre va énoncer l'ensemble des antécédents du patient au début de chaque compte rendu, d'autres vont seulement mettre les informations permettant d'identifier le patient. Les centres n'ont pas les mêmes types de documents. L'objectif principal de cette étude est de définir à l'aide d'experts les différentes catégories à laquelle pourrait appartenir une phrase issue de ces documents. Cette analyse préliminaire effectuée conjointement avec les médecins a permis d'établir la liste des catégories qui permet d'extraire les informations de manière optimisée. Les catégories identifiées et leurs descriptions sont présentées dans le Tableau 5.5 ci-dessous.

TABLE 5.5 – Description des catégories identifiées pour la tâche de classification automatique des phrases.

Catégorie	**Description**
Antécédent personnel	Les antécédents personnels du patients sont l'ensemble des informations médicales thérapeutiques et faits antérieurs à une maladie, permettant de comprendre celle-ci et de juger de la conduite à tenir.
Antécédent familial	Les antécédents familiaux sont des informations médicales qui concernent les proches du patient et qui peuvent interférer avec les décisions de prescription d'examens complémentaires ou de traitements.
Entête	Les informations à faible valeur médicale et qui intéressent peu les médecins lors de la prise de décision médicale. Par exemple : toute information à caractère administratif telle que les noms, les adresses, etc.

Négation	Les phrases négatives servent à nier, à refuser ou à interdire quelque chose. Une phrase est considérée comme négative, si elle contient une forme de négation sur une partie ou toute la phrase.
Incertitude	Les phrases incertaines sont les phrases qui expriment un doute. La lecture seule d'une phrase incertaine ne permet pas d'affirmer/infirmer tous les faits qui y sont évoqués.
Métastase	Une phrase métastatique contient des informations relatives à l'existence d'une forme de métastase pour un patient.

Le but de l'identification des phrases négatives et incertaines est d'exclure les concepts médicaux extraits qui sont contenus dans les phrases. De cette manière, une compréhension sémantique cohérente de ces relations peut être garantie. Il est à noter qu'une phrase peut ne pas appartenir à aucune des catégories définies ci-dessus.

5.4.3 Analyse des travaux connexes

Avant le succès qu'ont connu les approches d'apprentissage automatique profond, une grande variété d'algorithmes souvent basés sur le lexique (tels que : Les arbres de décision (DT) (Morgan and Sonquist, 1963), les machines à vecteurs de support (SVM) (Vapnik, 1964), les k plus proches voisins (kNN) (Patrick and II, 1969), etc.) a été proposée pour résoudre le problème de classification de textes. Cependant, la structure sémantique entre les mots d'un texte a un rôle prépondérant dans la tâche de classification. L'utilisation de méthodes lexicales augmente la perte de relations sémantiques entre les mots d'un texte en négligeant le contexte de signification et/ou l'ordre des mots.

Avec l'émergence d'approches d'apprentissage profond, les architectures des réseaux de neurones récurrents (RNN) (Sutskever et al., 2011; Mandic and Chambers, 2001) et des réseaux de neurones convolutionnels (CNN) (Dos Santos and Gatti de Bayser, 2014; Wang et al., 2017) ont pu présenté de meilleurs résultats comparés aux algorithmes d'apprentissage classique, et ont permis ainsi d'améliorer les tâches de traitement automatique du langage naturel (Kowsari et al., 2019). Ces approches peuvent extraire automatiquement les caractéristiques essentielles d'un texte et modéliser les relations complexes et non linéaires qui peuvent exister dans les données (Schuster and Paliwal, 1997; Liu et al., 2013). Dans les RNN, les informations stockées dans les nœuds des couches précédentes sont prises en compte. Ce mécanisme permet de valoriser le contexte textuel dans lequel se trouve un mot et donne un sens à l'ordre des mots dans une séquence lors de la classification des textes, et facilite ainsi l'analyse sémantique d'un texte. Cependant, le calcul de la dernière valeur prédite dépend de nombreuses valeurs passées. LSTM (Hochreiter and Schmidhuber, 1997) est un type spécifique de RNN conçu précisément pour résoudre le problème de la disparition du gradient qui peut survenir lors de l'apprentissage du modèle en contrôlant la quantité d'informations autorisées dans chaque nœud (Nowak et al., 2017). Bien qu'elles soient efficaces,

ces approches présentent certains inconvénients. Une des limitations des LSTM est qu'il est nécessaire de lire une séquence entière pour produire une prédiction, ce qui se traduit par des performances plus lentes lorsqu'il s'agit de textes longs (Hochreiter and Schmidhuber, 1997).

Par ailleurs, initialement conçus pour le traitement et la classification d'images, les réseaux CNN (LeCun et al., 1989) ont donné des résultats très satisfaisants lorsqu'ils sont appliqués aux tâches du TALN (Chowdhary, 2020). Pour la classification de textes, plusieurs études, utilisant des CNN, montrent que sur la base des caractéristiques au niveau des mots et des caractères, une meilleure analyse sémantique des textes par rapport aux RNN est garantie (Phuong and Le, 2018; Shakya et al., 2018). Les CNN peuvent déterminer efficacement les parties les plus discriminantes du texte. Cependant, ces réseaux utilisent souvent la convolution avec une taille de fenêtre fixe sur les mots d'une phrase pour éviter de manipuler une quantité considérable de paramètres, ce qui entraîne une perte d'informations précieuses.

Dans cette section, nous allons proposer un modèle de classification des phrases issues de textes cliniques basée sur un réseau de neurones convolutionnel (CNN) combiné à un réseau de neurone récurrent (LSTM). Les vecteurs de mots utilisés en entrée du réseau ont été générés à l'aide d'un modèle de plongement de mots en utilisant *FastText*. Ce modèle bénéficie à la fois des avantages des CNN pour l'extraction des caractéristiques locales et aussi des dépendances à long terme capturées grâce à la capacité mémoire élevée des réseaux LSTM pour connecter correctement les entités extraites, assurant ainsi une meilleure précision lors de la classification des textes.

Bien que l'apprentissage du modèle soit très important, le principal problème pour l'utilisation des méthodes d'apprentissage réside dans la constitution d'un ensemble de données annotées représentatif du corpus.

5.4.4 Annotation des phrases

Les algorithmes d'apprentissage automatique s'appliquent à des données étiquetées (c'est-à-dire annotées) qui constituent l'ensemble d'apprentissage. L'annotation des phrases permet donc de constituer le jeu d'entraînement, et cela en attribuant à chaque phrase une étiquette significative (une catégorie parmi celles prédéfinies), en fonction de son contenu. La qualité de l'annotation est très importante pour la formation d'un modèle de classification pleinement généralisable. Les erreurs de mauvaises affectations peuvent corrompre l'apprentissage d'un modèle. En plus de la qualité des annotations, ce processus doit garantir l'exhaustivité afin que les phrases annotées soient représentatives du corpus global.

L'annotation de textes cliniques est une tâche difficile pour plusieurs raisons. D'une part, elle constitue une opération fastidieuse et très chronophage qui nécessite beaucoup de ressources et de connaissances métiers liées au domaine d'application. D'autre part, en raison de l'utilisation de termes spécifiques au domaine, de la temporalité variable exprimée dans les données et des particularités des textes traités (Botsis et al., 2010).

En dehors du domaine biomédical, des interfaces d'annotation à usage général ont été développées pour les tâches du TALN, telles que la classification ou la reconnaissance d'entités nommées (Islamaj et al., 2020; Yang et al., 2017). Il existe dans la littérature de nombreux outils, tels que BRAT (Stenetorp et al., 2012), qui permettent également de gérer la distribution, le suivi et la collecte des corpus. D'autres outils sont basés sur l'apprentissage comme DUALIST (Settles, 2011). Bien que ces outils soient matures et offrent des fonctionnalités avancées, ils peuvent être complexes à mettre en place et à utiliser. Les travaux antérieurs sur les textes cliniques comprennent, par exemple, MetaMAP (Aronson, 2001) et cTAKES (Savova et al., 2010). Ces deux outils ont fourni des interfaces pour inspecter les entités reconnues, mais ils ne permettent pas de corriger ou de modifier les concepts, ou de spécifier des annotations supplémentaires.

Dans le cadre de nos travaux, nous avons utilisé l'outil *Prodigy* (Montani and Honnibal, 2018) pour effectuer la tâche d'annotation. Cet outil permet d'annoter facilement les phrases afin de constituer un jeu de données annotées. Il dispose de trois modes d'annotation :

— Mode manuel.
— Mode semi-manuel, utilisant un modèle d'apprentissage pour suggérer une annotation.
— Mode automatique, utilisant un modèle pour demander l'avis de l'utilisateur uniquement sur les phrases où celui-ci est incertain.

Les modes manuel et semi-manuel permettent de sélectionner les classes appropriées pour une phrase donnée tandis que le mode automatique permet uniquement de valider ou de rejeter la proposition du modèle.

Pour constituer le jeu d'entraînement, des comptes rendus de différents types ont été aléatoirement extraits de notre corpus constitué de 360.000 comptes rendus. Ces documents ont été ensuite nettoyés, tokenisés et découpés en phrases. L'annotation ainsi que la formation des modèles de classification automatique sont des processus très itératifs. L'ajout de phrases bien annotées au jeu d'apprentissage permet d'augmenter les performances de classification. Le nombre total des phrases annotées est décrit dans le Tableau 5.6 ci-dessous.

TABLE 5.6 – Nombre de phrases annotées par catégorie.

Catégorie	**#Phrase**	**Pourcentage**
Antécédent personnel	3660	08.32 %
Antécédent familial	3453	07.86 %
Entête	5196	11.83 %
Négation	5180	11.79 %
Incertitude	6487	14.77 %
Métastase	7574	17.24 %
Autre	12368	28.16 %
Total	**43918**	**100.0 %**

Un total de 51.604 phrases dont 43.918 distinctes ont été annotées. Avec une moyenne de 36, 8 phrases par compte rendu. Chaque phrase se compose d'une moyenne de 25 *tokens*, avec un intervalle allant de 1 à 293 *tokens*.

5.4.5 Approche proposée

L'approche de classification proposée est composée de quatre couches : une couche d'entrée pour le plongements de mots (cf. section 5.3), une couche de neurones convolutive unidimensionnelle pour l'extraction d'entités locales, une couche de neurones LSTM pour capturer les dépendances à long terme et une couche de classification pour la prédiction des catégories précédemment définies. L'architecture du modèle proposé est illustrée dans la Figure 5.3 suivante.

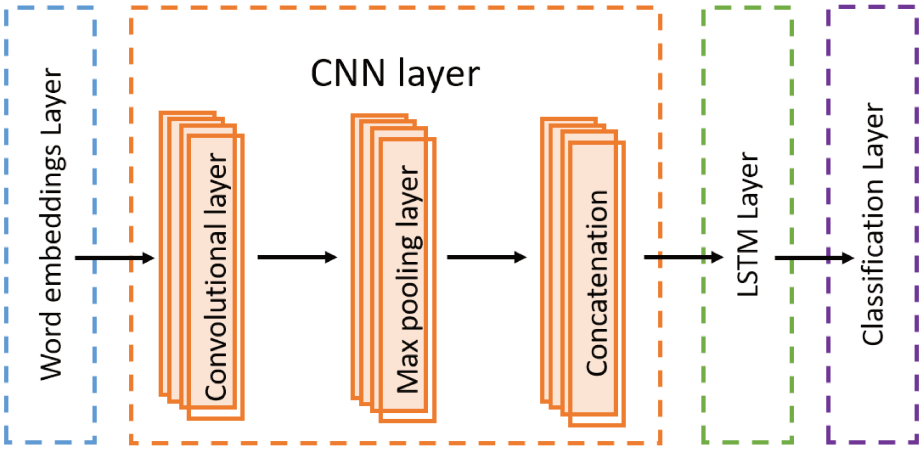


FIGURE 5.3 – Architecture proposée du modèle de classification automatique de phrases.

5.4.5.1 Couche convolutive à une dimension

Dans la tâche de classification, chaque phrase X_n est représentée par une séquence de n mots comme suit :

$$X_{1:n} = x_1 \oplus x_2 \oplus ... \oplus x_n \tag{5.2}$$

où $\oplus$ est l'opérateur de concaténation, et x_i représente le vecteur de mot à k dimensions du $i_{ième}$ mot de la phrase. Cette phrase est ensuite transformée en matrice, où chacune de ses colonnes correspond aux vecteurs des mots qui la composent. Cette matrice constitue le point d'entrée dans la couche de convolution. Une couche de convolution unidimensionnelle (Conv1D) est utilisée afin de capturer les informations séquentielles et réduire les dimensions des données d'entrée. Une opération de convolution implique un noyau convolutif appliqué à une fenêtre fixe $W \in R_k^n$ de mots pour calculer une nouvelle fonctionnalité. Ce noyau, également appelé filtre, complète l'extraction des fonctionnalités. Chaque filtre est appliqué à une fenêtre de m mots

pour obtenir une seule caractéristique. Pour garantir l'intégrité du mot comme la plus petite granularité, la largeur du filtre est égale à la largeur de la matrice d'origine. Dans la couche de convolution, une opération matrice-vecteur est appliquée à chaque fenêtre en utilisant la matrice de poids W pour obtenir une carte de caractéristiques $C \in R_{n-m+1}$. Le $i_{ième}$ élément de cette carte est défini par :

$$C_i = \sigma(\sum W \cdot [C_{i:i+m-1}]) + b) \tag{5.3}$$

Où b est le terme de biais utilisé pour ajuster la sortie ainsi que la somme pondérée des entrées du neurone. Cela permet de décaler la fonction d'activation non linéaire $ReLU$ notée ici σ. L'utilisation de $ReLU$ permet de réduire le nombre d'itérations nécessaires à la convergence dans les réseaux profonds. La même matrice est utilisée pour extraire les caractéristiques locales pour chaque fenêtre calculée de la séquence de mots d'entrée, extrayant ainsi le vecteur de caractéristiques n-grammes de taille $n - m + 1$. Nous utilisons plusieurs longueurs de filtre pour obtenir des caractéristiques variées et suffisantes. Ensuite, le résultat de la convolution est regroupé à l'aide de l'opération de mise en commun maximum (*max pooling*) pour capturer les caractéristiques essentielles du texte. Cette opération est souvent considérée comme une sélection de fonctionnalités lorsqu'il s'agit de traitement du langage naturel et elle offre de meilleurs résultats comparés à la mise en commun moyenne (*average pooling*), notamment en complexité de calcul. À l'issue de cette étape et pour améliorer la qualité de notre tâche de classification de texte, les différentes entités calculées sont concaténées pour constituer l'entrée de la couche LSTM.

5.4.5.2 Couche LSTM

Les RNN (Rumelhart et al., 1985) peuvent gérer l'entrée des informations séquentielles, et ont fourni d'excellents résultats dans plusieurs tâches du TALN (Chowdhary, 2020), mais il est difficile de former un RNN à capturer les dépendances à long terme car le gradient a tendance à exploser lorsque les séquences d'entrée sont longues. Dans l'approche proposée, les phrases sont analysées mot par mot, tout en préservant la sémantique de tout le texte précédant le mot courant dans une couche cachée. Par conséquent, le gradient peut ne pas être en mesure de se rétropropager (*backpropagation*) lorsqu'il traite du texte long en raison de nombreuses transformations de non-linéarité. Ce problème est la première motivation derrière l'utilisation du LSTM (Hochreiter and Schmidhuber, 1997) dans notre approche. LSTM permet de capturer efficacement des informations contextuelles à partir du texte d'entrée grâce à sa grande puissance de mémoire. Un tél réseau repose sur un mécanisme de porte qui lui permet de séparer les informations importantes à court terme en utilisant l'état caché des informations importantes à long terme en passant par l'état de la cellule. Une couche LSTM se compose de trois portes : (i) la porte d'oubli, qui décide quelles informations retirer de l'état de la cellule, (ii) la porte d'entrée, qui décide des informations à mettre à jour à partir de l'état de la cellule, et (iii) la porte de sortie, qui décide de la sortie finale du réseau. La cellule LSTM est représentée par la Figure 5.4.

Les informations sont transmises via deux canaux d'une cellule à l'autre, h et c.

1. À l'instant t, la relation de récurrence est calculée comme suit :

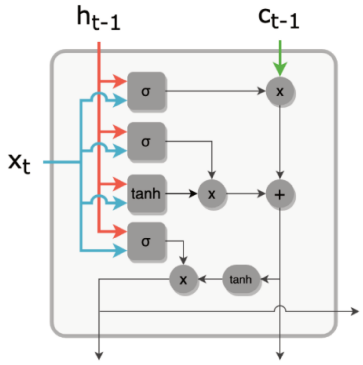


FIGURE 5.4 – Représentation d'une couche de réseau LSTM.

$$h_t, c_t = f(x_t, h_{t-1}, c_{t-1}) \tag{5.4}$$

Où x_t représente le mot courant de la séquence, h_{t-1} l'état caché de la cellule précédente et c_{t-1} l'état de cellule précédente. Ce dernier vecteur évite le problème du gradient de fuite car il est mis à jour de manière additive à chaque étape, sans passer par une fonction d'activation.

2. La porte d'oubli est une couche dense avec une activation sigmoïde qui agit comme un filtre pour oublier certaines informations sur l'état des cellules. A partir de h_{t-1} et x_t, cette porte d'oubli produit un vecteur :

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{5.5}$$

3. De même, la porte d'entrée produit un filtre i_t à partir de h_{t-1} et x_t.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{5.6}$$

4. En même temps, un vecteur g_t est créé par une fonction $tanh$, qui permet de mettre à jour l'état de la cellule.

$$g_t = \tanh(W_g \cdot [h_{t-1}, x_t] + b_g) \tag{5.7}$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes g_t \tag{5.8}$$

5. Similaire à f_t et i_t, le port de sortie produit un filtre o_t.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5.9}$$

6. Les valeurs du nouvel état de la cellule c_t sont réduites à l'intervalle $]-1, 1[$ par la fonction d'activation tanh. Un filtrage à travers la porte de sortie o_t est alors effectué pour obtenir finalement la sortie h_t.

$$h_t = o_t \otimes tanh(c_t) \tag{5.10}$$

Où σ désigne la fonction logistique sigmoïde, $\otimes$ désigne le produit vectoriel par élément, W sont les matrices de poids et b représente le terme de biais.

5.4.5.3 Couche de classification

Le dernier composant de notre modèle est une couche entièrement connectée (*fully connected layer*), qui prend en entrée les caractéristiques générées à partir d'une phrase par la couche LSTM, puis prédit la catégorie la plus appropriée en fonction du contenu sémantique et syntaxique. La probabilité qu'une phrase appartienne à une catégorie est calculée par la fonction d'activation *softmax*, comme suit :

$$P_i = \frac{exp^{o_i}}{\sum_{j=1} exp^{o_j}} \tag{5.11}$$

Où P_i indique la probabilité de la $i_{ème}$ catégorie, exp^{o_i} signifie la valeur correspondante de la sortie de la $i_{ème}$ catégorie, et j indique le nombre total de catégories.

5.4.6 Expérimentations et résultats

Afin d'évaluer les performances de classification du modèle proposé et de valider l'approche adoptée, plusieurs expérimentations ont été réalisées. Pour chaque catégorie identifiée, un gold standard a été constitué. Les phrases des gold standard ont été annotées avec *Prodigy* par les médecins. Contrairement à un ensemble d'entraînement qui peut être constitué pour simplifier la tâche de classification du modèle, le gold standard doit être annoté par un expert intransigeant afin de ne pas corrompre certaines annotations. Le modèle a été testé et comparé sur la tâche de classification de phrases à plusieurs approches de la littérature d'apprentissage automatique (arbres de décision (DT), Naïve Bayes (NB), machine à vecteurs de support (SVM) et les k plus proches voisins (kNN)), et d'apprentissage profond. Ces méthodes sont efficaces et ont obtenu d'excellents résultats dans la classification de texte. Chaque algorithme a été utilisé pour entraîner plusieurs modèles en faisant varier les paramètres d'entrée à chaque exécution (recherche gloutonne), afin d'identifier le modèle qui minimise le mieux la fonction de coût prédéfinie et ainsi pouvoir comparer les performances des meilleurs modèles de classification obtenus. Le Tableau 5.7 décrit les paramètres qui ont permis d'obtenir les meilleurs performances de classification.

TABLE 5.7 – Paramètres des algorithmes d'apprentissage automatique utilisés pour la classification des phrases.

Algorithme	**Paramètres**
Arbre de Décision	Extraction de caractéristique : TF-IDF. Alpha = 1 (Laplace smoothing).
Naïve Bayes	Extraction de caractéristique : TF-IDF. Alpha = 1. Profondeur maximale = 1.
Machine à Vecteurs de Support	Extraction de caractéristique : TF-IDF. Nombre d'itération = 1. Pénalité C = 1.

	Kernel = Poly. Tolérance pour le critère d'arrêt = 10^{-2}.
k Plus Proches Voisins	Extraction de caractéristique : TF-IDF. Nombre de voisins = 10.

De la même manière, différentes combinaisons de paramètres ont été expérimentées pour optimiser le modèle proposé. Les configurations suivantes fournissent le modèle le plus performant en termes de classification des phrases : (i) pas d'apprentissage : $0,01$, (ii) le nombre d'époques entre 15 et 20, (iii) *dropout* à $0,2$, (iv) méthode d'optimisation Adam, (v) le nombre de noyaux de convolution est de 256, et (vi) largeur des filtres multiples est (3, 4 et 5).

Pour évaluer les performances des différents modèles, les mesures utilisées comme critères d'évaluation sont : la précision, le rappel et le F-score. La précision mesure la capacité d'un modèle à classer uniquement les documents pertinents (précision), le rappel mesure la capacité d'un modèle à classer tous les documents pertinents (sensibilité) et le F-score est une moyenne harmonique calculée à partir de la précision et du rappel. Ces mesures sont largement utilisées dans l'apprentissage automatique et déterminent l'adéquation d'un algorithme de classification. Elles sont calculées à l'aide des formules suivantes :

$$Précision = \frac{Nombre\ de\ documents\ pertinents\ sélectionnés}{Nombre\ total\ de\ documents\ pertinents} \tag{5.12}$$

$$Rappel = \frac{Nombre\ de\ documents\ pertinents\ sélectionnés}{Nombre\ total\ de\ documents\ sélectionnés} \tag{5.13}$$

$$F - score = \frac{2 \times Précision \times Rappel}{Précision + Rappel} \tag{5.14}$$

Les F-score obtenues de chaque modèle sont présentées dans le Tableau 5.8 ci-dessous.

TABLE 5.8 – Nombre de phrases annotées par catégorie.

Catégorie	**DT**	**NB**	**SVN**	***k*NN**	**Modèle proposé**
Antécédent personnel	88,54%	87,01 %	91,13 %	90,33 %	**91,61 %**
Antécédent familial	90,51 %	86,37 %	90,37 %	90,54 %	**93,17 %**
Entête	87,65 %	85,46 %	87,91 %	86,19 %	**89,21 %**
Négation	94,84 %	92,55 %	96,57 %	95,03 %	**97,79 %**
Incertitude	83,72 %	80,67 %	86,41 %	86,44 %	**87,26 %**
Métastase	80,98 %	75,61 %	82,42 %	83,08 %	**88,13 %**
Total	86,83 %	83,55 %	88,34 %	87,90 %	**90,67 %**

Comme cela a été discuté dans la section 5.3, l'utilisation du même processus

de tokenisation pour l'entraînement des vecteurs des mots avec *FastText* et lors de la détection des limites des phrases permet l'obtention de résultats cohérents et une classification de phrases plus performante. De plus, l'utilisation de la couche LSTM dans le modèle proposé permet aussi d'améliorer les performances. Dans cette optique, les modèles les plus prometteurs ont été entraînés plusieurs fois, afin de garder le meilleur résultat. Les résultats (F-score) dans le Tableau 5.9 ci-dessous permettent de comprendre l'impact de chaque brique dans l'optimisation de la classification des phrases.

TABLE 5.9 – Nombre de phrases annotées par catégorie.

Catégorie	**FastText + CNN**	**Tokenisation + FastText + CNN**	**Modèle proposé**
Antécédent personnel	91,02 %	91,48 %	91,61 %
Antécédent familial	94,70 %	91,86 %	93,17 %
Entête	88,22 %	88,41 %	89,21 %
Négation	96,82 %	97,18 %	97,79 %
Incertitude	86,29 %	83,60 %	87,26 %
Métastase	83,55 %	87,47 %	88,13 %
Total	89,14 %	89,98 %	90,67 %

5.4.6.1 Discussion des résultats

Sur la base des valeurs de F-score obtenues, on peut voir que le modèle proposé obtient les meilleures performances de classification sur toutes les catégories, à l'exception des antécédents familiaux. Ces performances varient entre $87,26\%$ et $97,79\%$ respectivement pour les phrases incertaines et les phrases négatives. Premièrement, ces résultats confirment de manière générale que les modèles d'apprentissage profond permettent d'obtenir de meilleurs résultats comparés aux algorithmes classiques d'apprentissage automatique dans le domaine de la classification de textes. En effet, bien que les algorithmes Naïve Bayes (NB), les k plus proches voisins (kNN) et les machines à vecteurs de support (SVM) soient très efficaces, la grande dimensionnalité des caractéristiques d'un texte peut représenter particulièrement un frein à l'apprentissage du modèle. Par ailleurs, en dépit de la rapidité des algorithmes basés sur les arbres de décision (DT), ces derniers sont sensibles au bruit que les données peuvent contenir. Deuxièmement, on confirme une amélioration du F-score total suite à l'utilisation du processus de tokenisation avant le calcul des vecteurs de mots. Et enfin, on note que la combinaison d'un réseau de neurones convolutionnels (CNN) et avec un réseau de neurones récurrents (LSTM) améliore les performances de classification. Le réseau (CNN) extrait les caractéristiques locales de l'entrée, puis le réseau (LSTM), tel qu'il est proposé, caractérise mieux les informations sémantiques en fournissant des représentations des caractéristiques au niveau de la phrase. De cette manière, notre modèle tire parti des avantages à la fois du modèle CNN et du modèle LSTM, ce qui permet d'obtenir une précision de classification plus élevée que les autres modèles.

Il important de souligner le fait que les phrases annotées sont dans beaucoup de cas assez ambiguës, puisqu'elles appartiennent à un contexte très spécifique. À titre d'exemple, dans la catégorie des métastases, la présence de celles-ci est parfois sous-entendue dans la phrase et non indiquée de façon explicite par le mot «métastase». Au sens large, une métastase est une croissance cellulaire qui se produit à distance du site primaire de cette croissance et sans contact direct avec elle. On peut déduire de la phrase «*Présente des localisations cutanées et osseuses*» que l'on a une présence de métastase, car deux localisations différentes sont indiquées, ce qui suggère qu'une des deux fait référence à une métastase dérivée de la localisation primaire, alors que la phrase «*Présente des localisations cutanées*» ne permet pas de déduire de façon certaine que l'on parle d'une métastase et non d'une tumeur. Le cas des métastases est d'autant plus difficile qu'il existe de nombreux cas particuliers dans lesquels on peut affirmer une présence ou non. Les médecins ne considèrent pas une métastase dans le cas d'un «*ganglion métastatique*», par exemple, car celui-ci n'a pas forcément atteint un organe du corps du patient. À l'inverse différents facteurs peuvent indiquer une présence de métastase de façon implicite. Ce type d'ambiguïté existe aussi pour les autres catégories, même si elles traitent des domaines plus généraux, d'autant plus que l'annotation se fait par phrase sans connaître le contexte du reste du document.

5.5 Détection de la portée de négation et de l'incertitude

Comme précisé dans la section 4.9, l'approche proposée pour la détection des négations et des incertitudes dans le texte est composée de deux phases. La première consiste en une classification de phrases sous forme d'une couche de raffinement pour ne cibler que les phrases concernées. La deuxième est la constitution d'un modèle d'extraction d'entités nommées afin d'identifier la portée des déclencheurs d'une négation ou d'une incertitude. Pour ce faire, toutes les phrases négatives et incertaines, annotées pour la tâche de classification des phrases, ont été annotées une deuxième fois pour la construction d'un modèle d'extraction d'entités nommées. Le processus d'annotation est simple et consiste à sélectionner le marqueur d'une négation ou d'une incertitude ainsi que la partie de la phrase impactée par ce dernier. Il est important de préciser que certains déclencheurs peuvent avoir plusieurs portées discontinues dans la même phrase. Le modèle a ensuite été entraîné. Le jeu de données est divisé en deux parties égales, une pour l'entraînement, une autre pour le test. Les résultats obtenus sont présentés dans le Tableau 5.10 ci-dessous.

TABLE 5.10 – Résultats de détection des portées de négations et des incertitudes.

Catégorie	**Hyperparamètres**	**Performances**
Négation	0.2 dropout 20 itérations	Précision : 87,34 % Rappel : 84,61 % F-score : 85,95 %
Incertitude	0.2 dropout 20 itérations	Précision : 75,08 % Rappel : 73,63 % F-score : 74,35 %

Les performances de telles tâches dépendent de plusieurs facteurs : (i) la qualité des annotations, même si la négation et l'incertitude relèvent de domaine général, il est parfois difficile et ambigu de déterminer avec précision leurs portées. L'identification de ces portées dépendent fortement de la compréhension de l'annotateur. L'ensemble des données d'entraînement et les gold standard sont donc discutables quant aux valeurs proposées lors de l'annotation. (ii) la quantité de données annotées : en effet, avoir un jeu de données annotées suffisamment grand permet une meilleure représentativité des données du corpus. (iii) la complexité de la tâche à résoudre : dans nos travaux, le nombre de phrases utilisées pour l'incertitude est supérieur au nombre de phrases négatives utilisées (cf. Tableau 5.6). Cependant, la détection de la portée de négation dans les phrases est plus performante comparée à la détection de la portée d'incertitude. Le modèle note une différence de plus de 10% sur les valeurs de F-score obtenues. Cela est dû au très grand nombre de variations possibles pour l'expression d'une incertitude dans la langue française (cf. Section (4.9)).

5.6 Conclusion

Dans ce chapitre, nous avons présenté notre approche de classification automatique des phrases extraites des comptes rendus médicaux. Cette opération est souvent effectuée en amont des systèmes de traitement de l'information et de fouille de textes. Elle permet de catégoriser les phrases en différentes thématiques selon leur contenu. Ce processus permet alors de guider l'extraction d'informations.

En effet, en fonction des catégories attribuées, certaines phrases portent plus de valeurs que d'autres en raison des informations qu'elles contiennent, et donc n'ont pas le même degré de fiabilité. L'identification d'un cancer dans une phrase métastatique ne doit pas être traitée de la même façon qu'une identification d'un cancer dans une phrase d'en-tête.

L'approche proposée est constituée essentiellement de deux étapes. La première consiste à entraîner les vecteurs de mots pour représenter les textes de façon à extraire le plus de caractéristiques possibles. Pour cela, nous avons démontré le besoin d'utilisation d'un corpus assez large pour un apprentissage performant, et l'intérêt d'unifier le processus de tokenisation pour l'ensemble des traitements effectués sur les textes. La seconde étape est le choix de l'algorithme de classification. Dans nos recherches, nous avons prouvé que les algorithmes d'apprentissage profond fournissent de meilleurs performances comparés aux algorithmes d'apprentissage classique sur nos données. En plus, la combinaison d'un réseau CNN et d'un réseau LSTM permet de surpasser les performances d'un réseau de neurones convolutionnel en termes de classification de phrase sur l'ensemble de nos données.

Toutefois, en dépit de l'importance du choix de l'algorithme de classification, la quantité et la qualité des données annotées, à passer en entrée des algorithmes, constitue un élément fondamental dans la classification automatique de données. L'un des points clés, pour bien réaliser cette tâche est de pouvoir disposer d'un jeu de données assez large pour des soucis de représentativité d'un corpus plus global, et d'une anno-

tation de qualité sans de mauvaises affectations qui pourraient corrompre le modèle d'apprentissage.

Dans le chapitre 6 suivant, nous allons discuter de la structuration des patients et de l'identification de concepts composites à l'aide des informations extraites. Ensuite, nous allons étudier la modélisation des réponses aux traitements et des phénomènes liés aux résistances et aux toxicités chez les patients cancéreux.

CHAPITRE 6

RÉSISTANCE

Sommaire

6.1 Introduction

Défini par une multitude de pathologies distinctes, le cancer est la première cause de décès en France (Boulat T, 2019). En dépit des progrès de la détection du cancer et de l'augmentation significative de l'arsenal thérapeutique, les phénomènes d'échappement et de résistance aux traitements restent des problématiques majeures conduisant chaque année aux décès de plusieurs millions de patients dans le monde et environ 150000 personnes en France. Dans ce contexte, l'objectif principal de cette thèse est l'identification et la caractérisation des patients résistants aux traitements d'oncologie en se basant sur les connaissances contenues dans les comptes rendus médicaux des dossiers patients. Le but de cette démarche est de fournir, in fine, la possibilité aux médecins d'étudier les similarités qui peuvent exister entre les populations des patients résistants à des fins thérapeutiques. Ce processus s'inscrit parfaitement dans une logique d'une médecine personnalisée. Afin de répondre à cette problématique, il est primordial d'étudier le processus de traitement des patients cancéreux et d'analyser les réactions de ces derniers aux traitements qui leur sont administrés.

Dans ce chapitre, nous allons présenter une modélisation de la résistance aux traitements d'oncologie réalisée à partir des connaissances extraites des comptes rendus. Pour cela, nous allons d'abord établir une structuration des patients afin de restituer leurs parcours palliatifs. Cette structuration dépend fortement de la compréhension des mécanismes de soins et de la qualité des informations extraites avec les méthodes présentées dans les chapitres 4 et 5. Cette étape est très importante et constitue l'un des éléments clés de notre approche. Nous allons ensuite définir nos méthodes de détection des réponses aux traitements et des toxicités, composantes fondamentales à la modélisation de la résistance. Pour chaque étape, nous allons présenter les expérimentations réalisées et discuter les résultats obtenus.

6.2 Quelques notions de biologie

6.2.1 De la cellule cancéreuse à la tumeur

Le cancer est une maladie connue par les médecins depuis plusieurs siècles. Cette maladie est provoquée par la transformation des cellules qui deviennent anormales et prolifèrent de façon excessive. La prolifération incontrôlée des cellules cancéreuses aboutira à la formation d'une tumeur maligne.

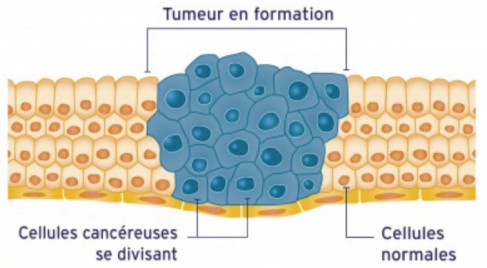


FIGURE 6.1 – Schéma représentant la formation d'une tumeur.

— Dès que la tumeur atteint 1 ou 2 millimètre(s), les cellules cancéreuses déclenchent l'angiogenèse, c'est-à-dire la formation de nouveaux vaisseaux sanguins qui irriguent et alimentent la tumeur. Sans cette irrigation sanguine, la tumeur ne pourrait continuer à se développer.
— Les cellules cancéreuses s'insinuent dans les tissus sains voisins et s'échappent de leur lieu d'origine pour développer des tumeurs secondaires (métastases) dans d'autres organes.
— Les cellules cancéreuses "détournent" à leur profit les cellules qui les entourent et les utilisent à leur avantage. Une tumeur est toujours formée par un agglomérat de cellules cancéreuses et de cellules normales qui collaborent entre elles.

De nombreux progrès dans le domaine de la médecine ont permis l'identification et la guérison de plusieurs de ces pathologies. Cependant, elle reste considérée comme une maladie incurable en raison des caractéristiques de la maladie auxquelles les médecins sont confrontés.

6.2.2 Étapes successives de l'évolution d'un cancer

En l'absence de traitement, la majorité des cancers (tumeurs dites « solides ») évolue en suivant les mêmes étapes, mais à des vitesses très variables et selon des modalités propres à chaque type de cancer. L'évolution du cancer est décrite plus en détail dans (Mesa, 2015; Mitchison et al., 2019). Le processus d'évolution naturelle de cette maladie en dehors de toute intervention ou de traitement est souvent répartie sur 4 phases :

1. Présence des lésions précancéreuses contenant des cellules en cours de transformation. Ces lésions n'engendrent pas forcément des vrais cancers.
2. Apparition et multiplication des cellules cancéreuses. La taille de la tumeur est petite et est attachée dans le tissu d'origine.
3. Grossissement de la tumeur et envahissement des tissus voisins.
4. Apparition des métastases d'abord dans les ganglions lymphatiques utiles dans la lutte contre les infections, ensuite dans les autres organes du corps.

Ce processus d'évolution n'est pas inéluctable. Il existe différents traitements et moyens plus ou moins lourds pour le patient afin de lutter contre la maladie, selon le type du cancer et le moment où la maladie est découverte : la chirurgie, la chimiothérapie, la radiothérapie, les thérapies ciblées, et la radiologie interventionnelle, etc. Une fois sous traitement, le schéma type d'un patient atteint du cancer suit en général le processus suivant :

1. Le patient réagit avec une réponse partielle au bout de quelques mois.
2. La maladie reste stable (par exemple pendant une année) : ce qui correspond à des observations reproductibles entre différents scanners.
3. Maladie évolutive : découverte de nouvelles évolutions des tumeurs. On retourne alors au premier point avec un arrêt, un changement ou une maintenance d'un traitement.

Ce dernier point est particulièrement intéressant à étudier et à analyser dans le cadre de notre recherche. Afin d'être en mesure de modéliser les résistances au traitement, il est essentiel de comprendre les mécanismes et les différents types et niveaux de réponse au traitement. Dès lors, plusieurs questions se posent naturellement à nous : à quel moment dans le parcours de soin d'un patient doit-on calculer la réponse ou la résistance au traitement ? L'arrêt d'un traitement est-il causé par une réponse complète, un changement protocolaire ou une toxicité ? etc. Pour répondre à ces questionnements, nous allons présenter dans un premier temps le modèle de structuration des patients atteints du cancer. En effet, dans le cadre de notre recherche, nous nous basons essentiellement sur des dossiers patients, notamment sur les connaissances contenues dans les comptes rendus médicaux. La bonne exploitation des ces dernières nous permettrait donc de répondre à notre problématique, à savoir la modélisation de la résistance chez les patients cancéreux. Pour ce faire, il est primordial de définir un modèle de données au sens informatique afin de faciliter l'extraction des informations d'une part, et de pouvoir en tirer le maximum de connaissances médicales quant à leur exploitation d'autre part. Dans ce qui suit, nous allons présenter notre méthode de structuration des patients.

6.3 Structuration des patients

La structuration d'un patient consiste à établir un modèle informatique de données dans lequel est regroupé toutes les informations pertinentes à caractère médical ou administratif extraites à partir des données lui appartenant qu'elles soient structurées ou pas. Comme nous l'avons évoqué dans les chapitres précédents, chaque dossier patient contient une centaine de documents hétérogènes. Cette diversité des sources utilisées permet de garantir et d'offrir à notre étude le moyen de prendre en considération le maximum d'informations nécessaires à la modélisation du phénomène de la résistance chez les patients cancéreux. Les comptes rendus médicaux représentent la grande majorité des documents patients, et sont représentés sous forme de textes non structurés. Compte tenu de la valeur indispensable des connaissances contenues dans ces derniers, il est tout à fait naturel de vouloir les extraire à des fins thérapeutiques ou de recherches.

L'identification et l'extraction de ces informations constituent un point de départ d'une longue chaîne de traitement. En effet, les connaissances extraites doivent être modélisées de manière intelligente afin de pouvoir répondre à une série de questionnements. Cette modélisation doit permettre de faciliter le travail des médecins au quotidien et permettre de répondre aux requêtes qu'ils peuvent formuler pour la recherche d'un patient ou la création d'une cohorte de patients homogènes dans le cadre de leur recherche.

Dans nos travaux, afin de pouvoir modéliser et étudier les résistances chez les patients cancéreux, nous avons présenté notre approche de structuration des patients. Cette étape est importante puisque, par définition, elle nous permet une extraction facile des concepts médicaux, notamment ceux qui ont un lien direct avec les différentes réponses aux traitements. En effet, la résistance est une forme de réponse à un traitement administré au patient, qui est souvent traduite par une évolution de la maladie, à

savoir la dernière étape du processus d'évolution de cette maladie. Pouvoir distinguer les différentes réponses entre elles est une étape critique pour la compréhension des résistances, en s'appuyant sur toutes les connaissances exploitables stockées dans les différents comptes rendus d'un patient. Notre approche est effectuée selon les deux étapes suivantes :

Dans un premier temps, une phase de structuration au niveau des documents est nécessaire pour valider ou discréditer les concepts identifiés en fonction du contexte dans lequel ils sont employés. En effet, les comptes rendus sont traités phrase par phrase. Or, nous avons démontré dans les précédents travaux que le contexte de la phrase n'est pas suffisant pour comprendre l'intégralité du document. De plus, les informations contenues dans certaines sections ont plus de valeur et de fiabilité que d'autres. Tous les concepts extraits au niveau des documents doivent être croisés pour l'identification des concepts composites. Ces derniers sont constitués généralement de plusieurs concepts élémentaires identifiés au niveau document. À titre d'exemple, les traitements sont créés à partir du regroupement des molécules de chimiothérapie et des protocoles, puis ils sont rattachés à une date si elle existe.

Dans un deuxième temps, une phase de structuration au niveau patient est effectuée, notamment pour la gestion des dates de début ou de fin d'événements tumoraux. En raison des informations dupliquées dans les différents comptes rendus d'un patient (rappel des antécédents médicaux, des symptômes, des diagnostics, etc), et le nombre important de documents générés par le suivi régulier des patients, il est par exemple difficile de déterminer avec précision la date de début d'un cancer ou d'une ligne de traitement. Cette deuxième phase de structuration permet de reconstituer le parcours de soin de chaque patient.

6.4 Modèles de données

Nous allons présenter dans ce qui suit deux modélisations simplifiées. La première au niveau document, et la seconde au niveau patient. Ces deux modèles de données représentent le fondement de notre méthode de modélisation et d'identification des résistances chez la patients cancéreux.

6.4.1 Modèle de données au niveau document

Afin de répondre aux questions les plus courantes dans un contexte de cancer, un modèle de données commun définissant la maladie cancéreuse a été défini, et permet de classer et d'organiser tous les concepts identifiés et extraits à partir des documents. Ce modèle repose sur plusieurs grandes classes hiérarchiques : les cancers (toutes les récidives de cancer pour un patient donné), les événements tumoraux (tumeur primaire, rechute locale ou métastatique), les actes (traitements et/ou analyses) et les documents (tous les documents d'un patient ou échantillons biologiques disponibles).

Le modèle de données présenté dans la Figure 6.2 montre les différents concepts (élémentaires ou composites) identifiés essentiellement à partir d'un compte rendu.

L'ensemble des documents structurés du patient sont utilisés à des fins d'enrichissement sémantique.

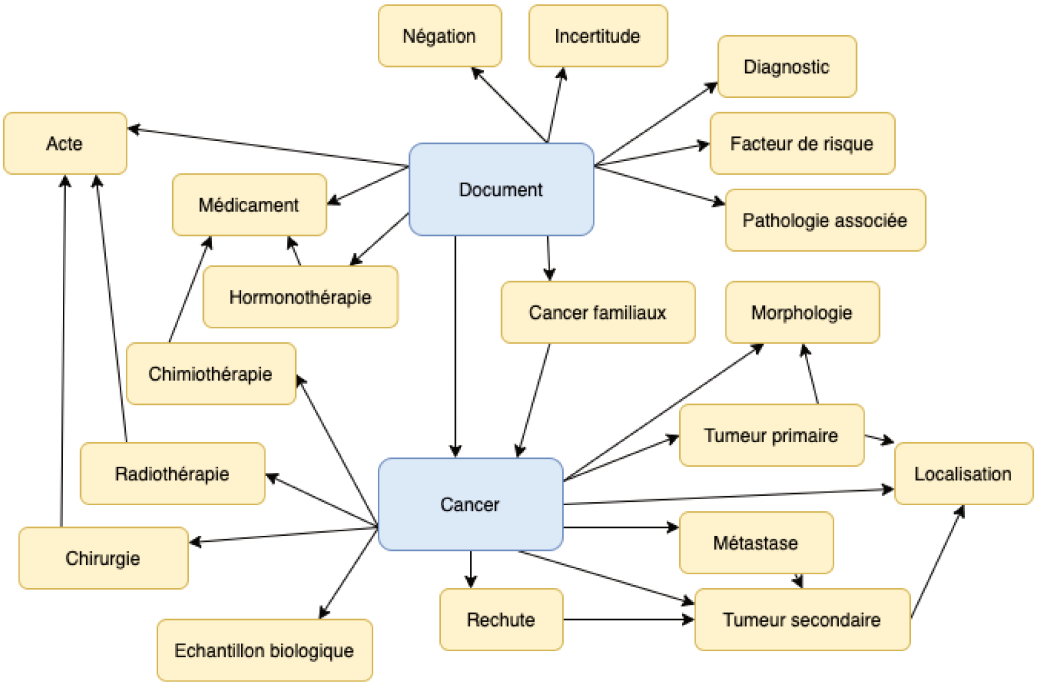


FIGURE 6.2 – Schéma du modèle de données niveau document.

6.4.2 Modèle de données au niveau patient

Les données du modèle document sont ensuite utilisées dans un mécanisme d'identification de certains événements tels que des rechutes locales ou métastatiques qui ne sont pas bien identifiés au niveau document. Ces événements sont donc représentés dans un modèle de données patient (Figure 6.3).

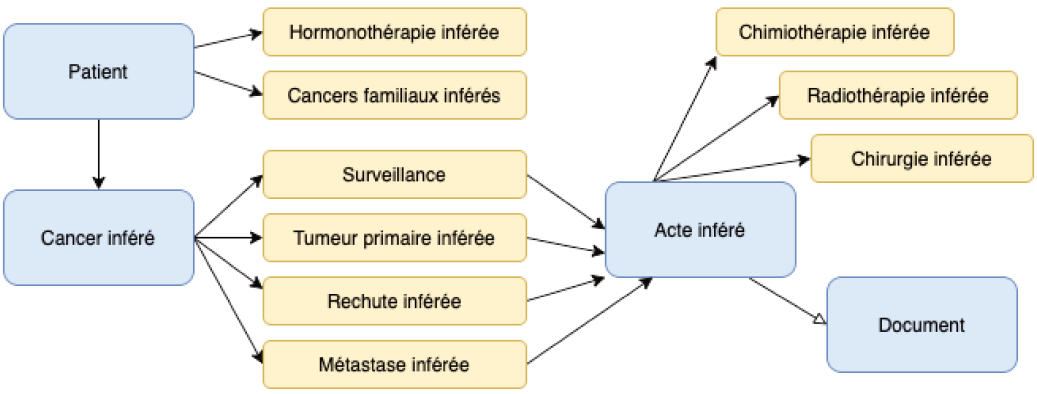


FIGURE 6.3 – Schéma du modèle de données niveau patient.

Le mécanisme d'inférence repose sur trois structures principales :

— Poids du document : les concepts identifiés dans un rapport de pathologie ont un poids plus important que ceux identifiés dans un rapport de consultation.
— Seuil d'inférence : un concept identifié seul dans un rapport ne déclenche pas la création d'un événement. Par exemple, un événement métastatique est structuré si et seulement s'il y a une répétition de termes associés dans un délai prédéfini.
— Gestion temporelle : le traitement de l'inférence fonctionne dans des délais spécifiques édictés par des règles médicales préalablement définies (ex : une rechute locale ne peut pas être déclenchée avant 90 jours après le dernier traitement subi).

Ces deux modèles nous permettent alors de fournir une représentation événementielle de la maladie carcinologique de chaque patient comme représentée dans la Figure 6.4 suivante.

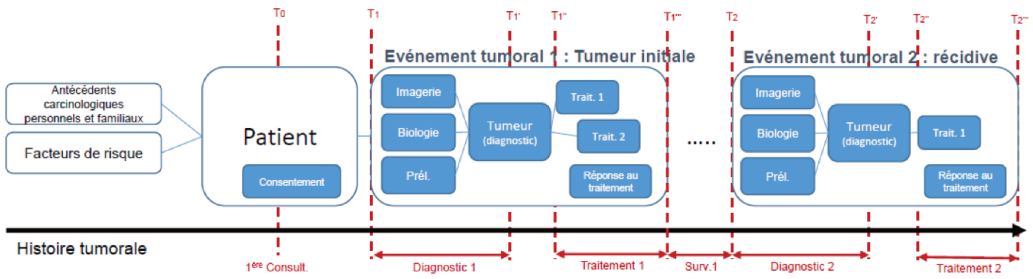


FIGURE 6.4 – Représentation événementielle de la maladie carcinologique d'un patient

Chaque patient appartenant au corpus des patients cancéreux a au cours de sa vie un ou plusieurs événements de type «*Cancer*». Chacun est composé d'un ou plusieurs événements de type «*Évolution Tumorale*». Durant chacun de ces événements, le patient subit un ou plusieurs traitements : événements de type «*Traitement*».

— Le concept *Patient* définit les caractéristiques de l'individu atteint d'un événement carcinologique et les informations relatives à son état.
— Le concept *Évènement Tumoral* définit l'ensemble des événements tumoraux survenant tout au long de la prise en charge d'un patient. Un évènement tumoral décrit aussi bien le cancer primaire, que la récidive loco-régionale ou métastatique.
— Le concept *Traitement* définit un traitement administré au patient au cours d'un événement tumoral donné. Ce traitement peut être de différents types : chirurgie, chimiothérapie, radiothérapie, immunothérapie.
— Le concept *Antécédent* recense l'ensemble des antécédents carcinologiques connus du patient et/ou de ses apparentés.
— Le concept *Facteur de risque* définit les causes primaires qui engendrent le développement d'un évènement tumoral, et peuvent être de nature différente : biomécanique, environnementale ou encore individuelle.

6.5 Modélisation des résistances au traitement

6.5.1 Définition du problème

La résistance aux traitements se traduit par l'inefficacité des thérapies suivies par les patients pour contrer la maladie. On distingue généralement deux types de résistances, à savoir la résistance primaire, où le traitement se révèle d'emblée inefficace, et la résistance secondaire ou acquise où la maladie se remet à progresser (évolution tumorale) après quelques semaines voire des mois de traitement. Dans ce dernier cas, le traitement n'est plus efficient pour le patient.

Les évolutions tumorales regroupent tout ce qui est tumeur primitive, récidive, métastase et la période durant laquelle le patient est sous surveillance. Ces différents évènements sont identifiés à partir des traitements extraits essentiellement des comptes rendus.

Les résistances aux traitements sont alors susceptibles d'apparaître pendant, ou à la suite, d'un traitement correspondant à une évolution tumorale. Plus généralement, on s'intéresse en particulier aux éventuels changements des lignes de traitements successives dans le parcours médical d'un patient, qui sont interprétés soit par une toxicité, une résistance ou simplement un changement conforme aux protocoles.

6.5.1.1 Signaux forts et signaux faibles

Un changement de ligne de traitement (Fin de traitement, décès du patient, etc.) est dû dans la majorité des cas (80% selon les médecins) à une résistance. De ce fait, cela représente un signal fort pour chercher à identifier les résistances. Cependant, d'autres évènements secondaires (signaux faibles) peuvent apporter de nouvelles informations utiles à la détection (ou non) des résistances, notamment, durant la période de surveillance, lors des soins palliatifs et après une pause puis reprise d'un même traitement. Ce dernier cas se produit par exemple lorsque le patient est sur le point d'effectuer une radiothérapie et donc doit justifier d'une pause (représente 10% des cas pour le cancer du pancréas).

TABLE 6.1 – Différents types de réponse au traitement.

Évènement principal (Signaux forts)	**Évènement secondaire (Signaux faibles)**	**Conclusion**
Arrêt de traitement	Surveillance	-
Arrêt de traitements systémiques généraux	Soins palliatifs	Résistance
Arrêt de traitements systémiques généraux	Pause thérapeutique et reprise du même traitement	-
Arrêt de traitements systémiques généraux	Radiothérapie et reprise du même traitement	-
Décès du patient	-	Résistance

Compte tenu des besoins de notre étude, l'identification des résistances pour les trois pathologies étudiées (Sein, Pancréas et Poumon) se fait dans le cas de diagnostic de métastases suivies d'un traitement systémique général. Il s'agit d'un traitement utilisant des substances qui se déplacent via le système sanguin et qui affectent les cellules au travers de tout le corps. Parmi ces traitements : la chimiothérapie conventionnelle, la thérapie ciblée, l'immunothérapie et l'hormonothérapie). Pour cela, nous allons chercher à identifier les résistances en se rapportant aux évolutions (évènements) tumorales et en particulier les métastases, et les fins de traitements. En effet, la résistance constitue la cause essentielle d'échec des traitements systémiques. Les résistances des patients aux traitements peuvent apparaître éventuellement à la fin d'une ligne de traitement. Celle-ci est initiée sur la base de changement de molécules. Le lancement d'une deuxième ligne de traitement peut être causé par une résistance, une toxicité, ou un changement de traitement protocolaire. En d'autres termes, le changement d'une ligne de traitement n'est pas nécessairement associé au phénomène de la résistance, mais il peut correspondre à un changement d'entretien qui fait partie du protocole établi, ou à l'apparition d'une toxicité.

Le schéma de la Figure 6.5 suivante permet d'illustrer les causes de changement de ligne de traitement.

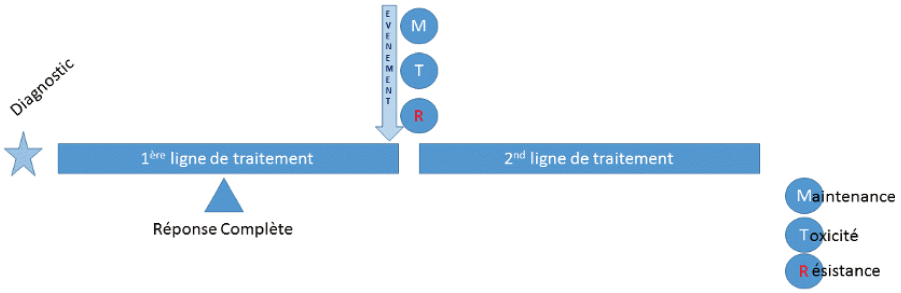


FIGURE 6.5 – Schéma du modèle de données niveau document.

La résistance est souvent considérée comme une progression de la maladie. Il s'agit de la troisième étape (Maladie évolutive) dans le processus d'évolution d'un patient cancéreux sous traitement (cf. 6.2.2). Cette progression est l'une des formes que peut prendre une réponse au traitement. L'identification de la cause d'arrêt ou de changement d'un traitement permet d'affirmer ou d'infirmer l'existence d'une éventuelle résistance au traitement. Dans un cadre plus général, l'idée c'est de détecter d'abord les différents niveaux de réponses, ensuite étudier les possibilités de résistance lorsqu'il s'agit de progressions. Dans un premier temps, nous allons nous focaliser sur l'identification des réponses au traitement à partir des comptes rendus.

6.5.2 Identification et détection des réponses au traitement

6.5.2.1 Définition du problème

La modélisation de la notion de réponse à partir du dossier patient est une problématique complexe, et demande dans un premier temps une simplification du problème.

L'évaluation de la réponse est faite systématiquement à chaque introduction de nouveaux éléments d'imagerie (scanner, données d'imagerie, etc). Les médecins demandent au patient d'effectuer de nouveaux examens quand ils ont des soupçons d'un changement dans son état de santé. Les médecins font ensuite des interventions adaptées à l'état de santé du patient en fonction des réponses aux traitements. Ces interventions peuvent être par exemple : l'arrêt d'un traitement, la mise sous surveillance, ou la mise sur soins palliatifs, etc.

Il existe plusieurs standards de mesure de réponse. Le plus connu sont les critères (*Response Evaluation Criteria In Solid Tumors*) (Kusaba and Saijo, 2000; Watanabe et al., 2009). Ces critères permettent de donner une indication indirecte de l'évolution de la tumeur. Selon ces critères, il existe plusieurs niveaux de réponse en fonction de la réaction des patients, de l'évolution des lésions cibles, et des examens effectués. Le Tableau 6.2 suivant résume les quatre niveaux de réponse en oncologie.

TABLE 6.2 – Les quatre niveaux de réponse au traitement en oncologie selon les critères RECIST.

Niveau de réponse	**Définition**
Réponse complète (RC)	Disparition de toutes les lésions. De plus, tous les ganglions lymphatiques (cible ou non-cible), doivent avoir atteint une dimension < 10 mm dans leur plus petit axe.
Réponse partielle (RP)	Une diminution de la somme des diamètres des lésions cibles, en prenant comme référence les diamètres de la somme de base ou une diminution du nombre de lésions.
Maladie stable (MS)	La tumeur reste à peu près de la même taille en prenant comme référence les plus petits diamètres, et aucune autre tumeur n'apparaît.
Progression (PR)	Augmentation de la somme des diamètres des lésions cibles par rapport à la plus petite somme des diamètres observée durant l'étude (*NADIR*), y compris la visite de baseline. En plus de cette augmentation, cette somme doit augmenter d'au moins 0,5 cm. L'apparition d'une ou plusieurs nouvelles lésions est également considérée comme progression.

Le dernier niveau de réponse dans le tableau «*Progression*», semble le plus proche de la notion de résistance au traitement.

6.5.2.2 Challenges et analyse de la solution

Une phase d'analyse a été d'abord effectuée, durant laquelle différents comptes rendus contenant au moins une réponse au traitement ont été étudiés. Si les différents niveaux de réponse sont clairement définis, leur identification dans les comptes rendus et interprétation est une tâche extrêmement complexe, et cela pour plusieurs raisons.

— Compte tenu de la complexité liée au domaine, lors de leur évaluation de la réponse, les médecins n'utilisent pas des formulations simples. Une maladie stable peut être énoncée de manière négative. Par exemple, dans la phrase «*Pas d'augmentation de la tumeur*», le médecin souhaite probablement souligner le caractère stable de la maladie. Il est donc nécessaire de gérer quelques formes de négations que nous avons identifiées dans les chapitres 4 et 5. Par ailleurs, des termes comme «*dégradation*» et «*progression*» bien qu'ils soient souvent utilisés dans des contextes opposés, ils peuvent tous deux exprimer une progression de la maladie.
— L'interprétation des réponses quand elles sont identifiées en se basant sur le vocabulaire est une tâche difficile. La phrase «*Diminution de 20% de la taille de la tumeur*», à l'encontre de ce qu'on pourrait imaginer, les médecins considèrent ici qu'il s'agit d'une maladie stable et non pas une réponse partielle.
— Dans les comptes rendus, il est souvent question de notions de réponses exprimées sans référence à la méthodologie de mesure précise utilisée et qualifiée de manière imprécise à l'aide d'adjectifs. En effet, les critères d'évaluation sont rarement cités de manière explicite. De plus, ces critères diffèrent selon les centres, les types de cancer traités et les types de réponses.
— Dans le cas d'un cancer avec plusieurs foyers, il est également possible d'avoir dans une même phrase des réponses partielles sur certains foyers, mais rien sur d'autres. Dans ce cas, le médecin peut parler d'une maladie stable, ou d'une réponse partielle. Plus rarement, on peut avoir une réponse partielle sur un foyer et une dégradation sur un autre.
— Les comptes rendus radiologiques ne sont pas nécessairement comparables. En effet, lors d'un examen, le radiologue dispose des anciens clichés. Cependant rien n'indique qu'il mesurera la même tumeur et/ou métastase. Ainsi, si le radiologue change entre deux examens, les mesures peuvent diverger.
— Enfin, en plus des différents niveaux de réponse, il existe plusieurs types de réponse exprimés dans les comptes rendus de consultation en fonction des examens réalisés (imagerie, anapathologie, PET-scan/scintigraphie, etc). Tous ces types ne sont pas nécessairement liés au phénomène de la résistance au traitement.

Ci-dessous quelques phrases exemples qu'utilisent les médecins pour évaluer la réponse au traitement.

- **progression** pulmonaire : traitement par TDM-1 (protocole KAMILLA).
- **Nette progression en taille** de la lésion hypermétabolique située sur le versant antérieur du dôme hépatique mesurée à environ 69 mm de grand axe transverse
- Aspect **stable** d'une légère hépatomégalie homogène.
- Au niveau du PET scan : on note une **réponse notable** puisqu'il y a une **régression** de l'ensemble des lésions hépatiques franches sur l'ensemble des lésions disséminées.
- **Nette régression** de l'épanchement péricardique qui paraît limité à la paroi latéro-VG.
- Cette **régression** peut être estimée à **+50 %**.
- Une **réponse métabolique complète** ganglionnaire sus-diaphragmatique, hépatique,

péritonéale et **complète ou quasi complète** ostéomédullaire.
- En revanche **augmentation** d'intensité de fixation de L2 et **apparition de 2 nouveaux foyers** hyperfixants suspects intéressant les parties postérieures des 2 ailes iliaques.
- IRM cervico-dorso-lombaire montre : la **stabilité** des lésions ossesuses.
- **Très nette majoration** de la lésion excavée lobaire supérieure gauche avec un comblement liquidien ou nécrotique central, épaississement de ses parois, et **majoration** de la condensation périphérique.
- IRM encéphalique : **réponse partielle** de la prise de contraste IRM (**-50%**) avec **disparition** de l'effet de masse.

Durant la phase d'analyse des comptes rendus, et avec l'aide des médecins, une liste des types de réponses a été établie. Le vocabulaire utilisé pour exprimer les réponses au traitement est partagé pour l'expression de tous les types de réponse. Pouvoir les distinguer permet alors d'accorder plus d'importance à certains types que d'autres. Le Tableau 6.3 suivant décrit les différents types de réponse au traitement identifiés.

TABLE 6.3 – Différents types de réponse au traitement.

Type de réponse	Description
Réponse d'état général	Concerne l'état de santé général du patient. Par exemple : l'état respiratoire ou neurologique.
Réponse métabolique	Évaluation de la réponse avec les critères *SUV*, (index utilisé pour caractériser la fixation du fluorodéoxyglucose en tomographie). Ce type de réponse est donné par PET-SCAN.
Réponse morphologique	Diminution considérable de la taille de la masse tumorale ou de son diamètre au scanner. Apparition de nouvelles lésions. Ce type de réponse est donné par l'imagerie (IRM, TDM, ...).
Réponse biologique	Nette diminution du taux des biomarqueurs tumoraux utilisés (pour le diagnostic ou la toxicité).

6.5.2.3 Approche proposée

Il existe essentiellement deux approches possibles pour la détection des réponses au traitement à partir des phrases des comptes rendus, soit mettre en place des règles de logique en se basant sur un dictionnaire de vocabulaire pour l'extraction des termes, soit entraîner un modèle de reconnaissance d'entités nommées à partir d'un ensemble de données annotées. Compte tenu des différentes raisons citées avant, l'approche proposée a été décomposée en deux étapes. La première étape consiste en un modèle NER pour l'identification des niveaux de réponse. La deuxième étape est réalisée à l'aide de règles pour identifier les types de réponses et les rattacher au niveaux de réponse détectés.

6.5.2.3.1 Détection des niveaux de réponse

Dans la phase d'analyse, il a été constaté que le vocabulaire utilisé pour évaluer une réponse est très riche et constitue une réelle source d'ambiguïté. Contrairement aux

règles, un modèle de NER permet de prendre en compte le contexte dans lequel une réponse est exprimée. Dans le but d'entraîner un modèle de reconnaissance d'entités nommées, il faut disposer d'un jeu d'entraînement. 2500 phrases contenant au moins un des quatre niveaux de réponse ont été extraites. Toutes ces phrases ont été ensuite annotées, en utilisant les différents niveaux de réponses comme des catégories.

Par ailleurs, la limite de séparation entre certaines de ces catégories n'est pas clairement définie, et dépend souvent de l'interprétation des médecins ou de la méthodologie de mesure utilisée. Par exemple, dans les phrases «*Bonne réponse partielle*» et «*Réponse partielle estimée à 20%*», un médecin pourrait identifier deux réponses partielles, alors qu'un autre médecin pourrait les dissocier en deux catégories différentes, une réponse partielle dans la première, et une maladie stable dans la deuxième. Cela représente un réel problème quant à l'annotation des données. Les termes désignant des «*progression*» sont souvent accompagnés d'adjectifs, par exemple :

- **Discrète progression** de la formation tumorale lobaire supérieure droite.
- Au niveau cervical, **très importante progression** de l'adénopathie décrite sur l'examen IRM précédent.
- Juin 2016 : **Nouvelle progression** sous la forme d'un lymphome folliculaire de bas grade.

Afin d'étudier de plus près ces ambiguïtés, et établir une limite claire entre une «*Réponse partielle*», une «*Maladie stable*» et une «*Progression*», les quatre niveaux de réponse ont été séparés en sept sous-niveaux. Le tableau 6.4 ci-dessous décrit les sous-niveaux identifiés, et présente le nombre d'entités annotées par catégorie.

TABLE 6.4 – Les sept sous-niveaux de réponse au traitement en oncologie

Sous-niveau de réponse	**Estimation**	**#Entités**
Réponse complète	-	143
Nette réponse partielle	$> 75\%$	386
Réponse partielle	entre 25% et 75%	594
Petite régression	$< 25\%$	129
Maladie stable	-	510
Petite progression	$< 25\%$	109
Progression	$> 25\%$	913
Non réponse	-	365

L'introduction de la catégorie «*Non réponse*» permet au modèle de reconnaissance d'entités nommées de comprendre ce qu'est réellement une réponse, et permet ainsi la gestion des cas particuliers. Quelques exemples sont présentés dans les phrases suivantes.

- Mastectomie partielle droite + GS : CCI 23 mm grade II, IM modéré.
- Dr Alt : informations données sur l'évolution de la maladie sous ENDOXAN.
- Reprise du CARBOPLATINE AUC 5 puis 4.5 sous traitement anti-allergique , 8 cures jusqu'au 04.06.2010.

- `Aggravation des infections urinaires basses.`
- `Présence de micronodules intra-hepatiques, compatibles avec des petites lésions métastatiques de tumeur endocrine.`

La constitution du jeu de données a été réalisée d'une manière itérative et incrémentale, pour intégrer le maximum de formules possibles quant à l'expression des réponses. Certaines phrases qui présentent des ambiguïtés ont été annotées séparément par deux annotateurs, puis validées par des médecins des centres de lutte contre le cancer.

6.5.2.3.2 Détection des types de réponse

L'identification des types de réponse permet de distinguer entre les réponses liées au phénomène de la résistance de celles qui ne le sont pas. En effet, après plusieurs séances de travail avec les médecins, seules les réponses au traitement de type «*morphologique*» sont intéressantes dans le cadre de l'identification des résistances. Bien qu'elle soit stricte, cette contrainte permet de décomplexifier la détection des réponses au traitement, et de réduire au maximum le bruit quitte à perdre des informations. Toutefois, il est important d'extraire tous les types de réponse.

Une fois les niveaux de réponse détectés, le contexte de la phrase est ensuite utilisé pour déterminer le type de chaque réponse à l'aide des règles et un vocabulaire construit conjointement avec les médecins lors de la phase d'analyse. Le référentiel des biomarqueurs est utilisé pour la détection des réponses de type «*biologique*». Cependant, le lexique n'est pas suffisant pour déterminer à lui seul le type de réponse approprié pour chaque niveau de réponse détecté. Pour surmonter ce problème, les concepts composites, détectés lors de la structuration des documents et des patients, sont exploités dans un moteur de règles.

1. $[Réponse] + [Évolution\ Tumorale] \Rightarrow [Réponse\ Morphologique]$.
2. $[Réponse] + [Localisation] \Rightarrow [Réponse\ Morphologique]$.
3. $[Réponse] + [Traitement\ chimio] \Rightarrow [Réponse\ Morphologique]$.
4. $[Réponse\ complète] \Rightarrow [Réponse\ Morphologique]$.

La dernière règle est appliquée si et seulement si toutes les autres règles ne sont pas applicables.

6.5.2.4 Résultats

Pour valider l'approche proposée pour la détection des niveaux de réponse, le modèle de reconnaissance d'entités nommées a été testé sur un gold standard annoté par les médecins du Centre Léon Bérard (CLB) contenant 589 entités de réponse au traitement. Cela permet d'apporter une compréhension neutre et une expertise dans le domaine de l'oncologie. Le Tableau 6.5 ci-dessous présente les résultats obtenus pour la détection des niveaux de réponse au traitement.

TABLE 6.5 – Résultats de la détection des niveaux de réponse au traitement.

Sous-niveau de réponse	Précision	Rappel	F-score
Réponse complète	100,0%	90,00%	94,74%
Nette réponse partielle	96,15%	96,15%	96,15%
Réponse partielle	81,82%	83,33%	82,57%
Petite régression	90,00%	69,23%	78,26%
Maladie stable	95,83%	85,19%	90,20%
Petite progression	100,0%	100,0%	100,0%
Progressions	97,59%	96,43%	97,01%

Plusieurs modèles ont été entraînés en faisant varier les paramètres d'entrée à chaque exécution (recherche gloutonne). Ainsi, dans le tableau ci-dessus ont été présentées les performances du meilleur modèle. Sur les 589 entités annotées ; l'approche proposée a obtenu une exactitude de $86,83\%$ sur la détection et le rattachement des types de réponse. Les réponses non détectées ou dont le type n'est pas bien identifié, sont pour la plupart causées par la non présence d'éléments de langage en mesure de fournir ces informations au niveau de la phrase.

6.5.3 Identification des toxicités

6.5.3.1 Définition du problème

Les traitements systémiques sont des méthodes thérapeutiques utilisant des médicaments toxiques pour la cellule cancéreuse, interférant avec son métabolisme ou sa division. L'augmentation de la dose de chimiothérapie par exemple entraîne toujours une toxicité accrue, s'accompagnant parfois, mais pas nécessairement, d'un effet thérapeutique supérieur. Ces traitements entraînent souvent divers effets secondaires perçus comme un problème médical inattendu qui survient pendant le traitement avec un médicament ou une autre thérapie. Ainsi, les patients ayant subi des traitements systémiques peuvent développer des toxicités qui provoquent l'arrêt ou le changement d'un traitement.

Après l'identification des réponses au traitement, la deuxième étape dans la détection des résistances est l'identification des toxicités à partir des comptes rendus. Les toxicités peuvent être une cause parmi d'autres d'arrêt ou de changement de traitement puisqu'elles sont, dans la plupart des cas, à l'origine de complications sévères. Leur identification permet alors d'infirmer l'existence d'une éventuelle résistance au traitement.

6.5.3.2 Challenges et analyse de la solution

La *CTCAE* (*Common Terminology Criteria for Adverse Events*, cf. Annexe A) est une terminologie souvent utilisée par les médecins dans les comptes rendus. Il s'agit d'une terminologie descriptive utilisée pour la déclaration des effets indésirables (des toxicités). Une échelle de grade (ou sévérité) allant de 1 à 5 est fournie pour chaque terme. Un tel effet est un signe, un symptôme ou une maladie non désirée, inattendue,

et associée chronologiquement à l'utilisation d'un traitement, ou d'une procédure. Un effet indésirable est un terme unique représentant un événement spécifique utilisé dans les comptes rendus et les analyses scientifiques. Le référentiel de la *CTCAE* contenant plusieurs milliers de termes (environ 400.000 termes), a été utilisé pour détecter les toxicités au niveau des phrases.

Une première phase d'analyse a été effectuée pour comprendre la nature des toxicités exprimées dans les comptes rendus médicaux, et tenir compte des cas particuliers qui peuvent exister. La difficulté principale réside dans l'ambiguïté de certaines phrases telles que :

- `11.2012 : épanchement pleural droit : TAXOL AVASTIN.`
- `FOLFOX Octobre 2005 - arrêt de l'OXALIPLATINE en raison de neuropathie, poursuite par LV5-FU2.`

Dans les deux exemples présentés ci-dessus il est montré qu'à la lecture seule des phrases, on n'est pas en mesure de trancher si le traitement a causé la toxicité, ou si cette dernière est soignée par le traitement. Il s'agit de phrases non marquées par des déclencheurs d'incertitude, ce qui rend l'élimination des concepts identifiés une tâche complexe voire impossible. De plus, certaines toxicités sont décrites sous la forme d'une définition dans le référentiel que l'on ne retrouve pas telle quelle dans les comptes rendus. Par ailleurs, l'exploitation des grades de toxicités est une tâche très complexe. Il n'y a pas de grade de toxicité clairement défini comme étant un motif suffisant pour le changement de ligne de traitement. Cela reste à l'appréciation des médecins lors de la période de surveillance des patients. Un dernier point important à souligner, est que les toxicités et les effets indésirables ne doivent pas nécessairement être causés par un médicament ou un traitement.

6.5.3.3 Approche proposée

L'objectif général de la détection des toxicités est de déterminer si elles sont ou non la cause d'un changement de traitement. Le référentiel *CTCAE* propose pour chaque toxicité de sa codification, un libellé ainsi que le détail de ses différents grades le cas échéant. Bien qu'il contienne nombre d'informations utiles, les libellés fournis ne sont alignés sur aucun autre référentiel, ce qui complexifie son utilisation. Dans notre recherche, les libellés des différentes toxicités ainsi que leurs identifiants associés ont été récupérés. Ces deux éléments ont permis de construire un référentiel de toxicités exploitable.

Compte tenu de différents challenges susmentionnés, l'approche de détection des toxicités est réalisée via la mise en place en amont d'un alignement entre le référentiel de toxicité *CTCAE* et le référentiel de diagnostics *CIM-10*. L'approche proposée s'articule autour des étapes suivantes :

1. **Première itération – Référentiels de toxicité :** Extraction des concepts *CIM-10* susceptibles d'être des toxicités quand ils sont liés à des traitements, pour les rapprocher du référentiel *CTCAE*. L'idée est ainsi de rapprocher ma-

nuellement les concepts les plus fréquents entre les deux référentiels. Cela a permis d'enrichir les concepts de toxicité.

2. **Seconde itération – Exploitation des référentiels :** Utilisation des référentiels précédents pour détecter les toxicités. La détection se fait à l'aide de règles (expressions régulières) pour qualifier ou non un diagnostic comme une toxicité.
3. **Amélioration continue :** La seconde itération a été reprise au fur et à mesure, tant pour améliorer les règles après application sur un jeu de données représentatif que pour optimiser le contenu des référentiels.

Par ailleurs, les termes les plus courants sont privilégiés puisqu'ils couvrent plus de 85% des cas de toxicités. Ce choix a été discuté avec les médecins pour se concentrer essentiellement sur les toxicités citées en lien avec un traitement systémique. Ainsi, 134 codes inusités en oncologie ont été écartés pour réduire au maximum la détection de faux positifs.

L'identification de la toxicité nécessite généralement un examen manuel des comptes rendus, car l'expertise humaine et la capacité de raisonnement sont nécessaires pour reconnaître les nuances des informations pertinentes contenues dans les comptes rendus. Pour créer un gold standard comme une base de test, un examen manuel des comptes rendus a été effectué par un médecin du centre *CLB*. Il a ensuite extrait 580 phrases contenant chacune au moins une toxicité avec des grades différents, qu'il a annoté manuellement. Sur un ensemble de 720 toxicités, cette approche a obtenu une exactitude automatique de $94,58\%$. Nous avons étudié alors le cas des toxicités non détectées. Dans tous les cas identifiés, il s'agit de formes ambiguës difficiles à comprendre même à la lecture de la phase.

6.5.4 Structuration des résistances au traitement

À présent, deux méthodes ont été mises en place pour l'identification et l'extraction des réponses au traitement et des toxicités à partir des comptes rendus (cf. 6.2). La modélisation des résistances passe par la structuration de ces concepts au niveau patient. Cette structuration a lieu une fois que tous les concepts ont été collectés à partir de tous les documents d'un patient. Cependant, il est important de souligner que la notion de résistance soit exclusivement liée au traitement. En d'autres termes, un même patient peut développer une résistance pour un traitement donné et être sensible pour un autre. Cela nous permet de cibler uniquement les patients ayant subi des traitements systémiques, et en particulier ceux qui ont au moins deux lignes de traitement. En effet, lors de la période de surveillance, à la détection d'une résistance les médecins changent systématiquement les traitements administrés au patient, et cela même si cette dernière est accompagnée d'une rémission complète ou d'une nette réponse partielle sur certains foyers. De ce fait, une deuxième contrainte stricte a été établie, et consiste à cibler les comptes rendus introduits dans une fenêtre de temps qui couvre la date de changement de traitement. Le positionnement de cette fenêtre temporelle a fait l'objet d'un processus itératif, avec une borne inférieure égale à trois mois avant la fin d'une ligne de traitement, et une borne supérieure égale à trois semaines après le début

de ligne de traitement suivante. Le Figure 6.6 suivante illustre cette fenêtre temporelle.

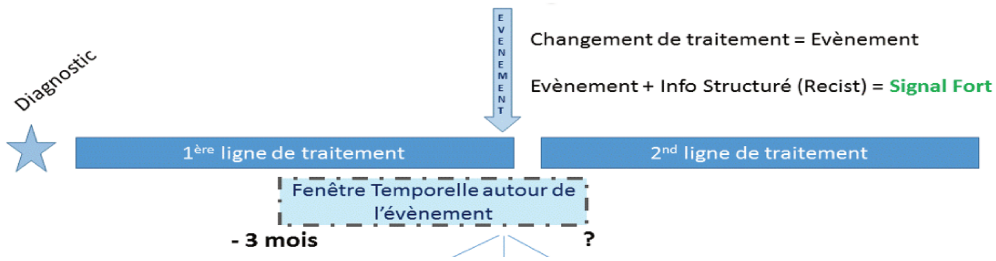


FIGURE 6.6 – Schéma du modèle de données niveau document.

Les concepts de réponses au traitement identifiées au niveau des documents sont utilisés dans un mécanisme de structuration des réponses au niveau patient en se basant sur un seuil d'inférence.

L'identification des réponses et des toxicités dans les comptes rendus introduits durant la période couverte par cette fenêtre temporelle doit permettre d'affirmer ou d'infirmer l'existence d'une éventuelle résistance au traitement avec le mécanisme suivant :

1. $[Toxicité] \Rightarrow [Infirmation\ de\ la\ Résistance]$.
2. $[Changement/arrêt\ protocolaire] \Rightarrow [Infirmation\ de\ la\ Résistance]$.
3. $[Progression\ morphologique\ la\ Maladie] \Rightarrow [Confirmation\ de\ la\ Résistance]$.
4. $[Réponse\ complète\ ou\ partielle] \Rightarrow [Infirmation\ de\ la\ Résistance]$.

Afin de tester l'approche proposée, les médecins du centre CLB ont sélectionné et annoté manuellement 132 patients qui ont développé des résistances au cours de leurs parcours de soins. La structuration des réponses au traitement a été réalisée plusieurs fois en variant la taille de la fenêtre temporelle. Cette approche a obtenu une précision de $35,24\%$, un rappel de $27,56\%$ et un F-score de $30,93\%$. Une étude approfondie a été ensuite conduite afin d'analyser les résultats obtenus. Plusieurs constatations ont été établies. Après une étude manuelle des dossiers patients, il s'est avéré que contrairement aux résistances, les progressions de type morphologique et les toxicités n'entraînent pas systématiquement un changement ou un arrêt de ligne de traitement. Ce premier constat est particulièrement contraignant, puisque la stratégie de modélisation des résistances au traitement est fondamentalement basée sur les changements de lignes de traitement. Sur l'ensemble de l'échantillon examiné, il existe une faible corrélation entre les progressions et les changements de lignes de traitements. De plus, ces derniers sont initiés sur la base de changement de molécules détectées à partir des comptes rendus. Cependant, rien ne nous garantit l'absence du bruit lors de leur détection. En effet, compte tenu de la nature non structurée des comptes rendus, et les particularités sémantiques de ces données, il est extrêmement difficile d'identifier

de manière sûre tous les concepts qui y sont contenus.

Par ailleurs, la résistance aux traitements anti-cancéreux est un phénomène extrêmement complexe, lié aux cellules tumorales et leur microenvironnement. À des fins de simplifications, les médecins nous ont proposé de cibler uniquement les phrases métastatiques pour la détection des réponses au traitement. Bien que cette simplification élimine beaucoup de bruit, elle reste une contrainte assez forte, et génère deux réels obstacles. Le premier obstacle réside dans la dépendance à la détection des phrases métastatiques. Or, le meilleur modèle de classification proposé a obtenu un F-score de $88,13\%$ pour la catégorisation des phrases métastatiques. Le second obstacle découle naturellement de cette simplification, puisque par définition, le périmètre de notre approche est en conséquence restreint, ce qui occasionne une perte d'information.

Bien que les connaissances contenues dans les comptes rendus soient incontournables pour l'amélioration des soins de santé de manière générale et d'aider les médecins lors de la prise des décisions médicales, ces documents sont insuffisants et parfois incomplets pour répondre à tous les questionnements des professionnels de santé. De plus, nous avons constaté qu'à la consultation manuelle des dossiers patients, les médecins peuvent avoir des avis divergents, voire opposés, quant à la confirmation ou l'infirmation des résistances au traitement. L'intégration des données d'imagerie et des données omiques, en particulier les données génomiques, serait alors une perspective parmi d'autres pour l'amélioration de la détection des résistances aux traitements. En effet, l'étude des données omiques permet le développement et l'application de nouvelles technologies pour la prévention de la maladie, notamment le cancer, en prenant en compte l'environnement auquel les cellules et/ou les protéines sont exposées, et de l'écosystème dans lequel elles interagissent.

6.6 Conclusion

Dans ce chapitre, nous avons présenté notre approche de détection des résistances aux traitement d'oncologie. Cette approche est basée essentiellement sur l'identification, l'extraction et la structuration des concepts des comptes rendus médicaux, et s'articule autour de deux points clés. La première problématique consiste à identifier les différents niveaux et types de réponse. Pour cela, nous avons mis en place un modèle de reconnaissance d'entités nommées pour la détection des niveaux de réponse. Ensuite, en se basant sur des règles et un vocabulaire de termes construits lors d'une phase d'analyse des comptes rendus, nous avons élaboré un mécanisme de détection et de rattachement des types de réponse aux niveaux de réponses appropriés. La seconde étape consiste à identifier les toxicités dans les comptes rendus. Pour ce faire, nous avons exploité les deux référentiels *CTCAE* et *CIM-10*, et présenté une méthode sur deux étapes. D'abord effectuer un alignement des deux référentiels pour en créer un référentiel de toxicités enrichi et adapté à notre problématique en cancérologie. Ensuite, utiliser des règles pour extraire et qualifier ou non un diagnostic comme une toxicité.

Finalement, une structuration des résistances aux traitements est réalisée en fonction de la nature des concepts identifiés précédemment. En particulier, ceux identifiés

dans les comptes rendus introduits durant la période de changement de traitement. Cela a pour but de confirmer ou d'infirmer l'existence d'une éventuelle résistance pour un traitement donné, puisque le changement de traitement est peut être protocolaire, causé par une toxicité ou par une résistance.

Testée sur un ensemble de 132 patients du centre *CLB*, cette approche a obtenu un F-score de $30,93\%$. Bien qu'ils soient loin d'être satisfaisants, ces résultats restent très encourageants. Afin de pallier le manque d'informations structurées et les limitations décrites dans la section précédente, l'intégration des données omiques et des données d'imagerie nous semble une bonne voie à étudier pour améliorer le score des détections des résistances aux traitements. Par ailleurs, la constitution d'un jeu de données plus riche permet sans doute d'approfondir les connaissances liées à ce phénomène et ainsi mieux comprendre les exigences qui en dépendent.

CHAPITRE 7

CONCLUSION

7.1 Bilan et contributions

Avec l'avènement de l'informatisation des données textuelles, plusieurs secteurs d'activité ont connu une forte révolution et une avancée sans précédent, notamment dans le monde de la médecine. Aujourd'hui, la médecine moderne est devenue presque inconcevable sans l'intégration des données numériques, qui ont bouleversé la compréhension scientifique des concepts élémentaires de la médecine, particulièrement avec la croissance exponentielle de la quantité de ces dernières. L'exploitation de ces données est devenue, de ce fait, plus que nécessaire afin d'aider et accompagner les professionnels de santé dans leur travail au quotidien, notamment dans la prise de décision médicale, et ainsi améliorer le processus de santé de manière générale. Cependant l'extraction de connaissances à partir des données médicales est un problème complexe. En effet, la plus grande partie de ces données se tient dans un format non structuré. De plus, la nature des informations quelquefois tacites qui y sont contenues nécessite un traitement au préalable afin de les comprendre et de les utiliser.

L'extraction d'informations pertinentes à partir des comptes rendus de dossiers patients est donc une tâche complexe liée à plusieurs sujets de recherche. Nous avons présenté dans cette thèse plusieurs contributions qui se situent dans le contexte général des systèmes d'extraction et de recherche d'informations, et plus précisément dans les techniques de traitement automatique de langage naturel. Par ailleurs, ces travaux s'inscrivent dans le cadre du projet *ConSoRe*, et ont pour but principal de répondre à la problématique de l'identification des patients résistants aux traitements d'oncologie à partir d'un corpus de données constitué essentiellement de comptes rendus médicaux enregistrés sous forme de textes libres non structurés.

Pour y parvenir, nous avons proposé des solutions aux différents problèmes liés à aux traitements des données médicales textuelles.

7.1.1 Pré-traitement des données

La première problématique que nous avons abordée est celle des pré-traitements des données textuelles. La complexité de cette tâche réside essentiellement dans la qualité

et la nature de ces textes. Compte tenu des particularités des comptes rendus médicaux traités (divergences typographiques, erreurs d'orthographe ou de syntaxe, etc) et de la spécificité du domaine médical, nous avons présenté une approche séquentielle de pré-traitement des données. En effet, outre les opérations de linguistiques standards telles que la tokenisation ou la segmentation en phrases, nous avons étudié plusieurs autres techniques dédiées au nettoyage des données ou encore à la correction orthographique par exemple.

Dans un premier temps, nous avons tout d'abord effectué une analyse manuelle des comptes rendus afin de comprendre la nature des informations véhiculées, et d'étudier les caractéristiques et les particularités des textes cliniques, et ainsi pourvoir répertorier les variantes des écritures utilisées dans le domaine de la cancérologie. Dès les premières constatations, nous avons réalisé l'absence d'une structure prédéfinie ou commune partagée par les comptes rendus, qui pourrait faciliter leur exploitation à posteriori. Par conséquent, nous avons proposé une approche de préparation des données textuelles composée d'une succession de méthodes appropriées à l'ensemble de ces données dans un processus unifié. Ces opérations de pré-traitement constituent d'ailleurs notre première contribution majeure, et visent à préparer les données aux tâches d'extraction et de recherche d'informations ultérieures.

Le processus de préparation des données textuelles que nous avons construit est constitué comme suit : (i) Collecte et nettoyage des données. L'intérêt étant de faciliter l'accès aux informations et de normaliser les données qui sont en l'occurrence issues de sources hétérogènes et enregistrées sous formats différents. Cette étape nous a permis d'éliminer les bruits et les parties non utiles tout en conservant toutes les caractéristiques textuelles essentielles pour les tâches ultérieures telles que la classification des phrases ou l'extraction d'entités nommées. Plusieurs transformations nécessaires ont été opérées telles que : l'extraction des textes brutes à partir des différents formats PDF, XML, MS Word, et HTML, ou encore la suppression des suites de caractères vides ou de lignes vides (inutiles). (ii) Une tokenisation afin d'identifier de manière juste les unités de traitements significatives, à savoir les mots (*token*). Cette opération est importante en particulier dans le traitement des données médicales où le vocabulaire et les termes métiers utilisés échappent souvent aux règles de la langue utilisée. En effet, afin de bien répondre à cette tâche, il était important de comprendre la nature des informations à traiter (le contenu et le style utilisés), mais surtout de garder à l'esprit les objectifs à atteindre derrière cette analyse de texte. En plus des règles strictes basées essentiellement sur les signes de ponctuation, nous avons intégré tous les référentiels (cf. Annexe A) utilisés lors de l'enrichissement sémantique pour bien identifier les mots (ou (*token*)) du jargon médical, et prendre en compte les différentes exceptions de tokenisation. (iii) Une méthode de détection des limites de phrases pour identifier la structure des documents traités et pouvoir les catégoriser par la suite selon leurs informations sémantiques. Cette opération consiste à transformer chaque compte rendu en une suite de phrases. Cela passe par l'identification des limites de phrases (début et fin). Cette segmentation constitue une étape clé dans notre approche notamment pour les tâches postérieures de classification des phrases et de reconnaissance d'entités nommées. Cette opération dépend en particulier des performances de la tokenisation

et s'appuie essentiellement sur les signes de ponctuation et la capitalisation des caractères. (iv) Une normalisation des textes à l'aide de la correction orthographique et des techniques de lemmatisation. En effet, comme illustré dans les exemples en Annexe B, les comptes rendus sont sujets à des fautes d'orthographe et de syntaxe qui peuvent corrompre les performances des tâches de classification. Pour cela, nous avons exploité une série de référentiels médicaux. (v) Une méthode de détection des dates et des événements temporels. Cette opération est essentielle et joue un rôle fondamental dans la reconstitution du parcours néoplasique des patients, puisqu'elle permet de dater les concepts médicaux des comptes rendus, et ainsi facilite la compréhension des informations extraites. La détection d'une date dans certaines phrases permet, par exemple, de les interpréter comme étant des phrases qui énoncent l'histoire de la maladie (cancer) chez un patient donné. L'identification de ces antécédents médicaux facilite donc aux médecins la prise de décisions médicales en choisissant des traitements plus appropriés pour le patient en question. (vi) Une méthode de désambiguïsation des termes tels que les protocoles, les métastases ou les codes *SNOMED-CT*. Pour la désambiguïsation des protocoles, nous avons entraîné des modèles d'apprentissage automatique pour la reconnaissance d'entités nommées. En revanche, pour la désambiguïsation des métastases, nous avons mis en place une méthode fondée sur la détection des termes ambigus à l'aide de règles à la fin du traitement du compte rendu. Enfin, (vii) Une méthode de détection des phrases négatives et incertaines afin d'exclure les concepts médicaux qu'on pourrait extraire lors de la recherche d'informations. En effet, la détection de telles phrases joue alors un rôle prépondérant dans les tâches de recherche et d'extraction d'informations, notamment dans le domaine médical où les médecins utilisent souvent la négation pour exclure un diagnostic ou un traitement, et des formulations hypothétiques sous forme d'incertitude afin de souligner la prudence avec laquelle ils souhaitent s'exprimer sur l'existence ou l'absence d'un événement tumoral.

7.1.2 Classification des données

La deuxième problématique que nous avons abordée est celle de l'extraction et de la recherche d'informations à partir des comptes rendus médicaux. À ce stade, les données ont subi une série de transformations et de pré-traitements afin de les préparer et de les rendre facilement accessibles. Notre deuxième contribution majeure consiste donc à extraire et faire valoir les informations pertinentes identifiées, et est répartie sur 4 axes. (i) L'extraction des concepts médicaux : nous avons utilisé plusieurs référentiels (cf. Annexe A) afin d'extraire et d'enrichir les concepts médicaux élémentaires, tels que les protocoles ou encore les biomarqueures, à partir des comptes rendus. Ces concepts vont être ensuite exploités pour reconstruire le parcours néoplasique pour chaque patient. Ce processus d'extraction de concepts est appliqué au niveau des phrases, et se base essentiellement sur des règles simples et des patterns qui permettent de détecter un concept (constitué d'un ou plusieurs *tokens*) dans un ensemble préalablement défini. (ii) La représentation des données et l'extraction des caractéristiques : afin de pouvoir exploiter nos données textuelles, nous avons procédé à une transformation des textes des comptes rendus en vecteurs numériques. Cette mathématisation est une étape importante souvent effectuée en amont de l'application des algorithme d'apprentissage automatique. Elle vise à représenter sous forme de vecteurs numériques les données textuelles afin de les rendre mathématiquement calculables. Dans l'approche que nous

avons présentée, nous avons opté pour les techniques de plongements de mots afin d'éviter les pertes d'informations sémantiques souvent causées par l'utilisation des modèles de sac de mots. Après l'étude de plusieurs algorithmes, nous avons entraîné un modèle *FastText* adapté particulièrement au contenu des nos comptes rendus où la présence des mots mal orthographiés n'est pas exclue. En effet, ce modèle prend en compte la similarité syntaxique des mots, de façon à ce que le vecteur d'un mot mal orthographié soit similaire (ou proche) dans l'espace vectoriel du mot sous sa forme correcte. (iii) La classification automatique des phrases : pour ce faire, nous avons mis en place un processus constitué de quatre étapes complémentaires à savoir : la représentation et l'extraction de caractéristiques, la réduction de dimensions, la sélection et la formation de classificateurs, et enfin l'évaluation de ses performances. Cette classification représente un moyen pour affiner l'extraction d'informations, et permet de choisir le traitement approprié à réaliser pour traiter les phrases. À titre d'exemple, l'identification des antécédents familiaux permet aux médecins de comprendre l'histoire de la maladie d'un patient et son contexte environnant. En revanche, à l'instar des phrases à caractère administratif, les concepts identifiés ne doivent pas s'inclure dans la recherche des pathologies associées à ce patient. L'approche que nous avons présenté est basée sur un modèle d'apprentissage automatique profond constitué d'une couche de plongement de mots, d'un réseau de neurones convolutionnel combiné à un réseau de neurones récurrent, et enfin une couche de classification entièrement connectée. Nous avons réalisé des expérimentations pour évaluer et valider les performances du modèle proposé et de les comparer aux performances des modèles entraînés à l'aide de différents algorithmes d'apprentissage automatique. (iv) La détection de la porte de négation et de l'incertitude : pour cette tâche, nous avons procédé en deux temps. D'abord une classification automatique des phrases qui constitue une couche de raffinement pour ne cibler que les phrases concernées (négatives ou incertaines), ensuite, la mise en place d'un modèle d'extraction d'entités nommées afin d'identifier la portée des déclencheurs d'une négation ou d'une incertitude.

7.1.3 Détéction des résistances

La troisième et dernière contribution majeure consiste en une approche originale de détection des résistances aux traitements en oncologie, en se basant exclusivement sur les informations extraites à partir de données médicales textuelles. Cette dernière contribution répond à l'objectif principal de notre thèse, à savoir l'identification et la caractérisation des patients résistants aux traitements d'oncologie en se basant sur les connaissances contenues dans les comptes rendus médicaux des dossiers patients. Pour cela, dans un premier temps, nous avons présenté deux modèles de structuration des données. La structuration des données d'un patient consiste à établir un modèle informatique de données dans lequel est regroupé toutes les connaissances pertinentes à caractère médical ou administratif extraites à partir des données lui appartenant. Le premier modèle de données que nous avons mis en place est introduit au niveau document et permet de structurer les informations identifiées dans ce dernier. Cette étape est nécessaire puisqu'elle nous permet de valider ou de discréditer les concepts identifiés dans un document en fonction du contexte dans lequel ils sont employés. En effet, les informations contenues dans certaines sections (notamment dans la fin des documents où les médecins ont tendance à lister les points importants sous forme de

conclusion) ont plus de valeur et de fiabilité que d'autres. Le second modèle de données est quant à lui introduit au niveau patient, et permet à partir des informations extraites dans plusieurs comptes rendus d'un même patient, reconstruire son parcours néoplasique. Pour y parvenir, nous avons défini un mécanisme d'inférence qui repose sur : le poids d'un document, un seuil d'inférence, et la gestion temporelle (par exemple les dates de début ou de fin d'événements tumoraux). En effet, les comptes rendus traités peuvent être de plusieurs types : prescriptions médicamenteuses, comptes rendus d'imagerie, comptes rendus de consultation, etc., et n'ont pas la même valeur puisque les informations qu'ils contiennent n'ont pas le même degré de fiabilité.

Dans un second temps, en se basant sur ces deux modèles, nous avons proposé deux approches respectivement pour la détection des réponses aux traitements et la détection des toxicités. La résistance à un traitement est l'inefficacité des thérapies suivies pour stopper la maladie, et est susceptible d'apparaître lors ou à la suite d'un traitement correspondant à une évolution tumorale (tumeur primitive, récidive, métastase). Dans nos travaux nous nous sommes focalisés en particulier sur les changements des lignes de traitements successives dans le parcours médical d'un patient, qui sont interprétés soit par une résistance, une toxicité ou simplement un changement conforme aux protocoles. En d'autres termes, le changement d'une ligne de traitement n'est pas nécessairement associé au phénomène de la résistance (même si cela couvre environ 80% des cas selon les médecins), mais il peut correspondre à un changement d'entretien qui fait partie du protocole établi, ou à l'apparition d'une toxicité. Nous avons donc mis une place une stratégie pour détecter les différents niveaux de réponse (RC, RP, MS et PR), et types de réponse (État général, métabolique, morphologique, ou biologique). Compte tenu de la nature des données que nous avons, l'approche proposée a été décomposée en deux étapes. La première étape consiste en un modèle NER pour l'identification des niveaux de réponse. La deuxième étape est réalisée à l'aide de règles pour identifier les types de réponses et les rattacher au niveaux de réponse détectés. Nous avons par ailleurs procédé à la détection des toxicités pour déterminer si elles sont ou non la cause d'un changement de traitement, et affirmer ou infirmer ainsi l'existence d'une éventuelle résistance au traitement. Pour cela, nous avons utilisé le référentiel CTCAE (cf. Annexe A).

Les concepts de résistance et de toxicité constituent des éléments indispensables quant à l'identification et la modélisation des résistances aux traitements. Pour répondre à cette problématique, nous avons finalement testé notre approche sur des ensembles de données annotées par les médecins des centres de lutte contre le cancer et présenté les résultats obtenus.

7.2 Perspectives

Nous proposons, dans le reste de ce manuscrit, les perspectives qui nous semblent essentielles, pour améliorer les résultats obtenus de nos travaux de recherche, et qui s'inscrivent parfaitement dans le cadre de notre projet, et ainsi pouvoir améliorer davantage notre approche de détection des résistances aux traitements anticancéreux.

Bien que les connaissances contenues dans les comptes rendus soient incontournables pour l'amélioration des soins de santé de manière générale et d'aider les médecins lors de la prise des décisions médicales, ces documents sont insuffisants pour répondre pleinement à une problématique complexe, à savoir l'identification des résistances aux traitements anticancéreux. En effet, ces données peuvent contenir des erreurs et/ou des manquements qui, bien qu'ils soient souvent marginales ou très peu généralisables, peuvent altérer l'exactitude des connaissances extraites, et ainsi peuvent corrompre les résultats obtenus. De plus, en dépit du poids inéluctable des informations contenues dans les comptes rendus, ces derniers ne représentent qu'une partie des données qu'on pourrait utiliser pour répondre à notre problématique. En effet, il existe d'autres types de données, par exemple les données d'imagerie médicales, qui peuvent nous apporter d'emblée des lectures différentes de notre problématique à travers leur exploitation, et nous mener à de sérieuses pistes de recherche. À partir de ce constat, il est de rigueur d'aller chercher et étudier d'autres sources d'informations quand cela est possible. Cette démarche nous permet non seulement de vérifier et valider les connaissances déjà acquises des comptes rendus, mais aussi de compléter l'arsenal que nous avons établi jusque là.

Dans les travaux futures, nous prévoyons d'intégrer de nouvelles sources de données, notamment les données omiques et les données d'imagerie (IRM, échographie, Scanner, etc). Ce processus permet d'une part d'apporter de nouvelles connaissances de valeurs sûres, et d'autres part de valider ou discréditer certaines informations issues des comptes rendus médicaux.

Dans la littérature, plusieurs travaux se sont focalisés sur l'utilisation des réseaux de neurones profonds pour accélérer le diagnostic par imagerie médicale. Il s'agit dans ces études de concevoir des modèles d'apprentissage automatique afin de segmenter les images de radiologie ou d'échographie par exemple, de manière à automatiser la détection des noyaux dans les cellules biologiques. À l'instar de ces travaux, notre première perspective consiste à adopter cette approche afin de surveiller l'évolution des cellules cancéreuses après chaque examen d'imagerie. Le but essentiel étant de détecter des marqueurs ou des indicateurs d'évolution de la maladie pour mieux comprendre les mécanismes de résistances chez les patients cancéreux.

Notre seconde perspective est de constituer de très larges ensembles de données annotées pour les différentes tâches de reconnaissances d'entités nommées et de classification de textes. Pour cela, il est important de solliciter les médecins et d'échanger davantage avec les professionnels de santé pour mieux comprendre la particularité et la diversité des connaissances contenues dans les données médicales. L'annotation de plus de données permet de construire des modèles d'apprentissage généralisables à l'ensemble de données stockées dans les systèmes d'informations des différents centres de la fédération UNICANCER. En revanche, cette opération est souvent fastidieuse et laborieuse. Il serait donc judicieux de considérer les techniques d'annotation semi-automatique ou automatique afin de réduire le temps consacré à l'annotation des données.

Cette deuxième perspective nous permettra dans un premier temps de concevoir des modèles d'apprentissage plus «intelligents» et plus «précis». En effet, même si l'optimisation des hyper-paramètres ne doit jamais être négligée afin de bien contrôler le processus d'apprentissage de manière générale, les chercheurs s'accordent à dire que l'annotation des données évaluée en termes de qualité et de quantité reste la clé essentielle pour l'amélioration des algorithmes d'apprentissage automatique profond.

Dans un second temps, le travail à mener dans cette perspective nous permet de publier un très large corpus de données annotées et le mettre à disposition des différents scientifiques qui pourraient à leur tour l'exploiter dans des travaux similaires ou complémentaires.

Bibliographie

H. Abdi and L. J. Williams. Principal component analysis. *WIREs Computational Statistics*, 2(4) :433–459, 2010.

H. Ahonen-Myka. Finding all maximal frequent sequences in text. *Proceedings of the 16th International Conference on Machine Learning ICML-99 Workshop on Machine Learning in Text Data Analysis*, pages 11–17, 1999a.

H. Ahonen-Myka. Knowledge discovery in documents by extracting frequent word sequences. *Library trends*, 48, 1999b.

H. Ahonen-Myka. Discovery of frequent word sequences in text. 2447 :180–189, 2002.

A. R. Aronson. Effective mapping of biomedical text to the umls metathesaurus : the metamap program. pages 17–21, 2001.

M. Attia. Arabic tokenization system. pages 65–72, 2007.

G. E. Bakx, L. Villodre, and G. Claramunt. Machine learning techniques for word sense disambiguation. *Unpublished doctoral dissertation, Universitat Politecnica de Catalunya*, 5, 2006.

G. Banks, H. Woznyj, R. Wesslen, and R. Ross. A review of best practice recommendations for text analysis in r (and a user-friendly app). *Journal of Business and Psychology*, 33 :445–459, 2018.

Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 5 :157–66, 1994.

D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel. Nymble : a high-performance learning name-finder. *arXiv preprint cmp-lg/9803003*, 1998.

D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3 :993–1022, 2003.

T. Botsis, G. Hartvigsen, F. Chen, and C. Weng. Secondary use of ehr : Data quality issues and informatics opportunities. *AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science*, 2010 :1–5, 2010.

M. C. F. L. R. S. G. R. Boulat T, Ghosn W. Principales évolutions de la mortalité par cause médicale sur la période 2000-2016 en france métropolitaine. *BEH. Bulletin épidémiologique hebdomadaire*, 29=30 :576–584, 2019.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2 :499–526, 2002.

R. Boyd. General use : Riot scan, 2016.

P. Bramsen, P. Deshpande, Y. K. Lee, and R. Barzilay. Finding temporal order in discharge summaries. In *AMIA annual symposium proceedings*, volume 2006, pages 81–86. American Medical Informatics Association, 2006.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification And Regression Trees*. 2017.

N. Béchet, M. Roche, and J. Chauché. Towards the selection of induced syntactic relations. volume 5478, pages 786–790, 2009.

M. Calvo-Flores, M. Martin-Bautista, D. Sánchez, and M. Vila. *Mining Text Data : Special Features and Patterns*, volume 2447, pages 175–186. 2002.

V. Canuel, B. Rance, P. Avillach, P. Degoulet, and A. Burgun. Translational research platforms integrating clinical and omics data : a review of publicly available solutions. *Briefings in bioinformatics*, 16(2) :280–290, 2015.

M. Chanchou, M. Reynier, A. Cougoul, J. Amat, B. Barrès, C. Bouvet, C. Valla, A. Kelly, C. Merlin, and F. Cachin. Valeur de la tep au 18fdg dans la prédiction de la réponse thérapeutique du cancer du sein her2+ : analyse intermédiaire. *Médecine Nucléaire*, 44 :92, 2020.

W. Chapman, W. Bridewell, P. Hanbury, G. Cooper, and B. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34 :301–310, 2001.

B. Y. M. Cheng, J. G. Carbonell, and J. Klein-Seetharaman. Protein classification based on text document classification techniques. *Proteins : Structure, Function, and Bioinformatics*, 58(4) :955–970, 2005.

K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv :1406.1078*, 2014.

K. Chowdhary. Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649, 2020.

J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv :1412.3555*, 2014.

K. Cohen and L. Hunter. Getting started in text mining. *PLoS computational biology*, 4 :e20, 2008.

T. Colin, F. Cornelis, J. Jouganous, M. Martin, and O. Saut. Patient specific image driven evaluation of the aggressiveness of metastases to the lung. volume 17, pages 553–60, 2014.

F. Colombani, E. Pereira, J. Bettaieb, L. Gobin, A. Cowppli-Bony, S. Hoppe, G. Coureau, M. Picat, R. Salamon, A. Monnereau, et al. Intérêt des données du registre hospitalier (enquête permanente cancer) d'un centre régional de lutte contre le cancer pour un registre de cancer en population. *Revue d'épidémiologie et de santé publique*, 61(1) :1–9, 2013.

G. V. Cormack and T. R. Lynam. Online supervised spam filter evaluation. *ACM Transactions on Information Systems (TOIS)*, 25(3) :11–es, 2007.

F. Cornelis, O. Saut, P. Cumsille, D. Lombardi, A. Iollo, J. Palussiere, and T. Colin. In vivo mathematical modeling of tumor growth from imaging data : Soon to come in the future ? *Diagnostic and interventional imaging*, 94(6) :593–600, 2013.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. Gate : A framework and graphical development environment for robust nlp tools and applications. pages 168–175, 2002.

H. Dalianis. *Clinical Text Retrieval - An Overview of Basic Building Blocks and Applications*, pages 147–165. Springer International Publishing, Cham, 2014.

C. Dalloux, V. Claveau, N. Grabar, L. E. S. Oliveira, C. M. C. Moro, Y. B. Gumiel, and D. R. Carvalho. Supervised learning for the detection of negation and of its scope in french and brazilian portuguese biomedical corpora. *Natural Language Engineering*, 27(2) :181–201, 2021.

J. Dawson. Suffix removal and word connation. *ALLC (Association for Literacy Linguistic Computing) Bulletin*, 2, 1974.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6) :391–407, 1990.

C. Dos Santos and M. Gatti de Bayser. Deep convolutional neural networks for sentiment analysis of short texts. pages 69–78), 2014.

M. Ducreux. Traitements oraux du cancer. *Soins*, 64 :20–21, 2019.

C. Duhamel, M. Bourgeau, G. Soto-Ares, F. Dubois, R. Assaker, E. Merlen, L. Mortier, M. Ilie, G. Raverot, and C. Cortet. Réponse dissociée d'un carcinome hypophysaire corticotrope après immunothérapie. *Annales d'Endocrinologie*, 81 :201, 2020.

A. Edwards. Is the reference in hartley (1749) to bayesian inference ? *American Statistician - AMER STATIST*, 40 :109–110, 1986.

W. Etaiwi and G. Al-Naymat. The impact of applying different preprocessing steps on review spam detection. *Procedia Computer Science*, 113 :273–279, 2017.

O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. Unsupervised named-entity extraction from the web : An experimental study. *Artificial Intelligence*, 165 :91–134, 2005.

F. Fancellu, A. Lopez, and B. Webber. Neural networks for negation scope detection. pages 495–504, 2016.

R. Feldman and H. Hirsh. Mining associations in text in the presence of background knowledge. In *KDD*, volume 96, pages 343–346, 1996.

R. Feldman, I. Dagan, and H. Hirsh. Mining text using keyword distributions. *J. Intell. Inf. Syst.*, 10 :281–300, 1998.

J. Gantz and E. Reinsel. Extracting value from chaos. *IDC's Digital Universe Study, sponsored by EMC*, 2011.

G. Giacinto and F. Roli. Methods for dynamic classifier selection. In *Proceedings 10th international conference on image analysis and processing*, pages 659–664. IEEE, 1999.

Y. Goldberg and O. Levy. word2vec explained : deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv :1402.3722*, 2014.

D. F. Gordon and M. Desjardins. Evaluation and selection of biases in machine learning. *Machine learning*, 20(1) :5–22, 1995.

A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks : the official journal of the International Neural Network Society*, 18 :602–10, 2005.

A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013.

G. Grefenstette and P. Tapanainen. What is a word, what is a sentence ? : problems of tokenisation. 1994.

N. Habash, O. Rambow, and R. Roth. Mada+tokan : A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*, page 62, 2009.

F. M. Hasan, N. UzZaman, and M. Khan. Comparison of different pos tagging techniques (n-gram, hmm and brill's tagger) for bangla. pages 121–126, 2007.

M. Hassler and G. Fliedl. Text preparation through extended tokenization. 2006.

L. A. Hazlehurst and W. S. Dalton. *De novo and acquired resistance to antitumor alkylating agents*, pages 377–389. Springer, 2006.

O. Heinonen, M. Klemettinen, and A. Verkamo. Finding co-occurring text phrases by combining sequence and frequent set discovery. pages 1–9, 1999.

D. Hindle. Acquiring disambiguation rules from text. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 118–125, 1989.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9 : 1735–80, 1997.

M. Hossin and M. N. Sulaiman. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2) :1, 2015.

A. Hotho, A. Nürnberger, and G. Paass. A brief survey of text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, 20 :19–62, 2005.

G. Hripcsak, N. Soulakis, L. Li, F. Morrison, A. Lai, N. Calman, and F. Mostashari. Syndromic surveillance using ambulatory electronic health records. *Journal of the American Medical Informatics Association : JAMIA*, 16 :354–61, 2009.

C.-R. Huang, P. Šimon, S.-K. Hsieh, and L. Prévot. Rethinking chinese word segmentation : Tokenization, character classification, or wordbreak identification. pages 69–72, 2007.

P. Irofti, A. Patrascu, and A. Băltoiu. *Fraud Detection in Networks*, pages 517–536. 2020.

R. Islamaj, D. Kwon, S. Kim, and Z. Lu. Teamtat : a collaborative text annotation tool. *Nucleic acids research*, 48(W1) :W5–W11, 2020.

H. Isozaki and H. Kazawa. Efficient support vector classifiers for named entity recognition. In *COLING 2002 : The 19th International Conference on Computational Linguistics*, 2002.

V. Ivančević, M. Knežević, M. Simić, D. Mandić, and I. Luković. Public healthcare and epidemiology with dr warehouse. 6 :329–342, 2013.

M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Reading text in the wild with convolutional neural networks. *International journal of computer vision*, 116 (1) :1–20, 2016.

A. G. Jivani et al. A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl*, 2(6) :1930–1938, 2011.

G. H. John and P. Langley. Estimating continuous distributions in bayesian classifiers. UAI'95, page 338–345, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

R. Johnson and T. Zhang. Effective use of word order for text categorization with convolutional neural networks. *arXiv preprint arXiv :1412.1058*, 2014.

I. Jolliffe. *Principal Component Analysis and Factor Analysis*, pages 115–128. 1986.

I. Jolliffe and J. Cadima. Principal component analysis : A review and recent developments. *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences*, 374 :20150202, 2016.

K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28 :11–21, 1972.

B. Jongejan and H. Dalianis. Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. pages 145–153, 2009.

H. Jongkind, G. Werff, and D. Batchelor. Patient administration management system (pams) a preliminary report. *European Journal of Cancer*, 29 :S258, 1993.

K. Jonsen, J. Fendt, and S. Point. Convincing qualitative research : What constitutes persuasive writing ? *Organizational Research Methods*, 21(1) :30–67, 2018.

A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov. Fasttext.zip : Compressing text classification models. *arXiv preprint arXiv :1612.03651*, 2016.

A. Kardan, F. Farahmandnia, and A. Omidvar. A novel approach for keyword extraction in learning objects using text mining and wordnet. *Global Journal of Information Technology*, 3 :6, 2013.

J. Kazama, T. Makino, Y. Ohta, and J. Tsujii. Tuning support vector machines for biomedical named entity recognition. pages 1–8, 2002.

M. L. Kern, G. Park, J. C. Eichstaedt, H. A. Schwartz, M. Sap, L. K. Smith, and L. H. Ungar. Gaining insights from social media language : Methodologies and challenges. *Psychological methods*, 21(4) :507, 2016.

A. Kilgarriff. "i don't believe in word senses". *Computers and the Humanities*, 31 : 91–113, 1997.

T. Kiss and J. Strunk. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32 :485–525, 2006.

V. Kobayashi, S. Mol, H. Berkers, G. Kismihók, and D. Den Hartog. Text classification for organizational researchers : A tutorial. *Organizational Research Methods*, 21 : 66–799, 2017.

V. B. Kobayashi, S. T. Mol, H. A. Berkers, G. Kismihók, and D. N. Den Hartog. Text mining in organizational research. *Organizational research methods*, 21(3) :733–765, 2018.

O. Kolak, W. Byrne, and P. Resnik. A generative probabilistic ocr model for nlp applications. pages 134–141, 2004.

K. Kowsari, D. E. Brown, M. Heidarysafa, K. J. Meimandi, M. S. Gerber, and L. E. Barnes. Hdltex : Hierarchical deep learning for text classification. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pages 364–371. IEEE, 2017.

K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown. Text classification algorithms : A survey, 2019.

K. Kukich. Techniques for automatically correcting words in text. *ACM Comput. Surv.*, 24 :377–439, 1992.

H. Kusaba and N. Saijo. A summary report of response evaluation criteria in solid tumors (recist criteria). *Gan to kagaku ryoho. Cancer chemotherapy*, 27 :1–5, 2000.

A. Labadié and V. Prince. Comparaison de méthodes lexicales et syntaxico-sémantiques dans la segmentation thématique de texte non supervisée. pages 330–339, 2008.

S. Lai, L. Xu, K. Liu, and J. Zhao. Recurrent convolutional neural networks for text classification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2267–2273. AAAI Press, 2015.

E. Lapponi, E. Velldal, L. Øvrelid, and J. Read. Uio 2 : sequence-labeling negation using dependency features. In ** SEM 2012 : The First Joint Conference on Lexical and Computational Semantics–Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 319–327, 2012.

L. S. Larkey. A patent search and classification system. pages 179–187, 1999.

Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Handwritten digit recognition with a back-propagation network. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1989.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86 :2278 – 2324, 1998.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521 :436–44, 2015.

C. Lee and G. G. Lee. Information gain and divergence-based feature selection for machine learning-based text categorization. *Information processing & management*, 42(1) :155–165, 2006.

R. I. Lerman and S. Yitzhaki. A note on the calculation and interpretation of the gini index. *Economics Letters*, 15(3-4) :363–368, 1984.

J. Lever, M. Krzywinski, and N. Altman. Points of significance : Classification evaluation. *Nature Methods*, 13 :603–604, 2016.

O. Levy and Y. Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 302–308, 2014.

D. D. Lewis. Feature selection and feature extraction for text categorization. In *Speech and Natural Language : Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

Y. Li and S. Chung. Text document clustering based on frequent word sequences. pages 293–294, 2005.

S.-H. Lin, C.-S. Shih, M. C. Chen, J.-M. Ho, M.-T. Ko, and Y.-M. Huang. Extracting classification knowledge of internet documents with mining term associations : a semantic approach. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 241–249, 1998.

C. Liu, W. Sun, W. Chao, and W. Che. Convolution neural network for relation extraction. volume 8347, pages 231–242, 2013.

Y. Liu, Y. Zhou, S. Wen, and C. Tang. A strategy on selecting performance metrics for classifier evaluation. *International Journal of Mobile Computing and Multimedia Communications (IJMCMC)*, 6(4) :20–35, 2014.

R. Lo, B. He, and I. Ounis. Automatically building a stopword list for an information retrieval system. *JDIM*, 3 :3–8, 2005.

J. B. Lovins. Development of a stemming algorithm. *Mech. Transl. Comput. Linguistics*, 11(1-2) :22–31, 1968.

J. Madec, G. Bouzillé, C. Riou, P. Van Hille, C. Merour, M.-L. Artigny, D. Delamarre, V. Raimbert, P. Lemordant, and M. Cuggia. *Studies in health technology and informatics*, 264 :1536–1537, 2019.

P. Majumder, M. Mitra, S. K. Parui, G. Kole, P. Mitra, and K. Datta. Yass : Yet another suffix stripper. *ACM transactions on information systems (TOIS)*, 25(4) : 18–es, 2007.

M. Makrehchi and M. S. Kamel. Automatic extraction of domain-specific stopwords from labeled documents. pages 222–233, 2008.

D. Mandic and J. Chambers. Recurrent neural networks for prediction : learning algorithms, architectures and stability. 2001.

B. Mathiak and S. Eckstein. Five steps to text mining in biomedical literature. In *Proceedings of the second European workshop on data mining and text mining in bioinformatics*, volume 24, pages 47–50. Citeseer, 2004.

A. McCallum and W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. 2003.

A. Meade and B. Craig. Identifying careless responses in survey data. *Psychological methods*, 17 :437–55, 2012.

K. Mesa. Essential cell biology. *The Yale Journal of Biology and Medicine*, 88 :100–101, 2015.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 2013.

N. Mitchison, A. Müllbacher, and K. Eichmann. *T-Cell Proliferation and Differentiation*, pages 171–176. 2019.

I. Montani and M. Honnibal. Prodigy : A new annotation tool for radically efficient machine teaching. *Artificial Intelligence*, to appear, 2018.

J. N. Morgan and J. A. Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, 58(302) :415–434, 1963.

Z. Muda, W. Mohamed, m. n. Sulaiman, and N. Udzir. K-means clustering and naive bayes classification for intrusion detection. *Journal of IT in Asia*, 4 :13–25, 2016.

S. Murphy, G. Weber, M. Mendis, V. Gainer, H. Chueh, S. Churchill, and I. Kohane. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association : JAMIA*, 17 :124–30, 2010.

P. G. Mutalik, A. Deshpande, and P. M. Nadkarni. Use of general-purpose negation detection to augment concept indexing of medical documents : a quantitative study using the umls. *Journal of the American Medical Informatics Association*, 8(6) : 598–609, 2001.

D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1) :3–26, 2007.

D. Nadeau, P. D. Turney, and S. Matwin. Unsupervised named-entity recognition : Generating gazetteers and resolving ambiguity. In *Conference of the Canadian society for computational studies of intelligence*, pages 266–277. Springer, 2006.

V. Nair and G. Hinton. Rectified linear units improve restricted boltzmann machines vinod nair. volume 27, pages 807–814, 2010.

I. Nancy and V. Jean. Word sense disambiguation : The state of the art. *Computational Linguistics*, 24(1) :1–40, 1998.

R. Navigli. Word sense disambiguation : A survey. *ACM computing surveys (CSUR)*, 41(2) :1–69, 2009.

J. Nowak, A. Taspinar, and R. Scherer. Lstm recurrent neural networks for short text and sentiment classification. pages 553–562, 2017.

W. Packard, E. Bender, J. Read, S. Oepen, and R. Dridan. Simple negation scope resolution through deep parsing : A semantic solution to a semantic problem. volume 1, pages 69–78, 2014.

C. Paice. Another stemmer. *SIGIR Forum*, 24 :56–61, 1990.

J. Paik, M. Mitra, S. Parui, and K. Järvelin. Gras : An effective and efficient stemming algorithm for information retrieval. *ACM Trans. Inf. Syst.*, 29 :19, 2011.

D. D. Palmer and M. A. Hearst. Adaptive multilingual sentence boundary disambiguation. *Computational linguistics*, 23(2) :241–267, 1997.

R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. pages 1310–1318, 2013.

E. Patrick and F. II. A generalization of the k-nearest neighbor rule. pages 63–64, 1969.

K. Paulson, V. Voillet, M. McAfee, D. Hunter, F. Wagener, M. Perdicchio, W. Valente, S. Koelle, C. Church, N. Vandeven, et al. Acquired cancer resistance to combination immunotherapy from transcriptional loss of class i hla. *Nature communications*, 9 (1) :1–10, 2018.

S. Paumier. *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. PhD thesis, 2003.

J. Pennebaker, M. Mehl, and K. Niederhoffer. Psychological aspects of natural language use : Our words, our selves. *Annual review of psychology*, 54 :547–77, 2003.

J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count : Liwc 2001. *Mahway : Lawrence Erlbaum Associates*, 71(2001) :2001, 2001.

J. Pennington, R. Socher, and C. Manning. Glove : Global vectors for word representation. volume 14, pages 1532–1543, 2014.

L.-H. Phuong and A.-C. Le. A comparative study of neural network models for sentence classification. pages 360–365, 2018.

J. Plisson, N. Lavrac, and D. Mladenić. A rule based approach to word lemmatization. 3 :83–86, 2004.

M. F. Porter. An algorithm for suffix stripping. *Program : Electronic Library and Information Systems*, 14, 1980.

R. Quinlan. Induction of decision trees. *Machine Learning*, 1 :81–106, 1986.

M. Rajman and R. Besançon. Text mining : natural language techniques and text mining applications. In *Data mining and reverse engineering*, pages 50–64. Springer, 1998.

J. Read, E. Velldal, L. Øvrelid, and S. Oepen. Uio1 : Constituent-based discriminative ranking for negation resolution. In ** SEM 2012 : The First Joint Conference on Lexical and Computational Semantics–Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 310–318, 2012.

M. Rogati, J. S. McCarley, and Y. Yang. Unsupervised learning of arabic stemming using a parallel corpus. pages 391–398, 2003.

A. Romaszewski, W. Trąbka, S. Jakubowski, M. Kielar, and Z. Kopański. Processing of medical data in the light of new technological and legal challenges. pages 27–32, 2019.

L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, and A. Verri. Are loss functions all the same ? *Neural computation*, 16 :1063–76, 2004.

S. Rose, D. Engel, N. Cramer, and W. Cowley. *Automatic Keyword Extraction from Individual Documents*, pages 1 – 20. 2010.

D. Rudrapal, A. Jamatia, K. Chakma, A. Das, and B. Gambäck. Sentence boundary detection for social media text. 2015.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

H. Saif, M. Fernandez, Y. He, and H. Alani. On stopwords, filtering and data sparsity for sentiment analysis of twitter. 2014.

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24 :513–523, 1988a.

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24 :513–523, 1988b.

G. Savova, J. Masanz, P. Ogren, J. Zheng, S. Sohn, K. Kipper-Schuler, and C. Chute. Mayo clinical text analysis and knowledge extraction system (ctakes) : Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association : JAMIA*, 17 :507–13, 2010.

D. Scherer, A. Müller, and S. Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. pages 92–101, 2010.

M. Schuster and K. Paliwal. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45 :2673 – 2681, 1997.

D. Schwab, J. Goulian, A. Tchechmedjiev, and H. Blanchon. Ant colony algorithm for the unsupervised word sense disambiguation of texts : Comparison and evaluation. In *Proceedings of COLING 2012*, pages 2389–2404, 2012.

H. A. Schwartz, J. C. Eichstaedt, L. Dziurzynski, M. L. Kern, E. Blanco, M. Kosinski, D. Stillwell, M. E. Seligman, and L. H. Ungar. Toward personality insights from language exploration in social media. In *2013 AAAI Spring Symposium Series*, 2013.

B. Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. pages 107–110, 2004.

B. Settles. Closing the loop : Fast, interactive semi-supervised annotation with queries on features and instances. pages 1467–1478, 2011.

S. Shakya, C. Zhang, and Z. Zhou. Comparative study of machine learning and deep learning architecture for human activity recognition using accelerometer data. 8 : 577–582, 2018.

C. Silva. The importance of stop word removal on recall values in text categorization. volume 3, pages 1661 – 1666 vol.3, 2003.

J. Simmons, L. Nelson, and U. Simonsohn. False-positive psychology : Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 20 :1–8, 2011.

V. Slavov, P. Rao, S. Paturi, T. Swami, M. Barnes, D. Rao, and R. Palvai. A new tool for sharing and querying of clinical documents modeled using hl7 version 3 standard. *Computer Methods and Programs in Biomedicine*, 112 :529–552, 2013.

R. Socher, J. Bauer, C. Manning, and N. Y. Parsing with compositional vector grammars. volume 1, pages 455–465, 2013a.

R. Socher, A. Perelygin, J. Wu, J. Chuang, C. Manning, A. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. *EMNLP*, 1631 :1631–1642, 2013b.

S. Sohail and E. Hassanain. Arabic email spam detection techniques and related arabic text preprocessing options : A survey. 2012.

A. B. Speer. Quantifying with words : An investigation of the validity of narrative-derived performance scores. *Personnel Psychology*, 71(3) :299–333, 2018.

P. Stenetorp, S. Pyysalo, G. Topic, T. Ohta, S. Ananiadou, and J. Tsujii. brat : a web-based tool for nlp-assisted text annotation. pages 102–107, 2012.

C. J. Stone. The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, pages 590–606, 1986.

D. Suhartono. Lemmatization technique in bahasa : Indonesian. *Journal of Software*, 9(5) :1203, 2014.

W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang. Data processing and text mining technologies on electronic medical records : A review. *Journal of Healthcare Engineering*, 2018 :1–9, 2018.

I. Sutskever, J. Martens, and G. Hinton. Generating text with recurrent neural networks. pages 1017–1024, 2011.

N. Syn, M. Teng, T. Mok, and R. Soo. De-novo and acquired resistance to immune checkpoint targeting. *The Lancet Oncology*, 18 :e731–e741, 2017.

M. D. Tapi Nzali, X. Tannier, and A. Névéol. Automatic extraction of time expressions accross domains in french narratives. pages 492–498, 2015.

S. Tata and J. M. Patel. Estimating the selectivity of tf-idf based cosine similarity predicates. *ACM Sigmod Record*, 36(2) :7–12, 2007.

D. Torunoğlu, E. Cakirman, M. Ganiz, S. Akyokus, and M. Z. Gürbüz. Analysis of preprocessing methods on classification of turkish texts. pages 112–117, 2011.

G. Trivedi, P. Pham, W. Chapman, R. Hwa, J. Wiebe, and H. Hochheiser. An interactive tool for natural language processing on clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 25, 2017.

A. A. Vapnik, V. et Chervonenkis. *A class of algorithms for pattern recognition learning*, volume 25, pages 937–945. 1964.

H. Wang and Y. Huang. Bondec–a sentence boundary detector. *CS224N Project, Stanford*, 2003.

J. Wang, Z. Wang, D. Zhang, and J. Yan. Combining knowledge with deep convolutional neural networks for short text classification. pages 2915–2921, 2017.

L. Wang, S. Li, D. F. Wong, and L. S. Chao. A joint chinese named entity recognition and disambiguation system. pages 146–151, 2012.

X. Wang, H. Zhang, X. Chen, and C. D. Resistance. Drug resistance and combating drug resistance in cancer. *Cancer Drug Resistance*, 2 :141–160, 2019.

Y. Wang. Various approaches in text pre-processing. 2004.

J. Wanous, S. Sullivan, and J. Malinak. The role of judgment calls in meta-analysis. *Journal of Applied Psychology*, 74 :259–264, 1989.

R. Warden. Impact of cabig on the european cancer community. *Ecancermedicalscience*, 5 :225, 2011.

H. Watanabe, M. Okada, Y. Kaji, M. Satouchi, Y. Sato, Y. Yamabe, H. Onaya, M. Endo, M. Sone, and Y. Arai. New response evaluation criteria in solid tumours - revised recist guideline (version 1.1). *Gan to kagaku ryoho. Cancer chemotherapy*, 36 :2495–501, 2009.

A. Weil. Precision medicine. *Health Affairs*, 37 :687–687, 2018.

A. Widlöcher and F. Bilhaut. La plate-forme linguastream : un environnement intégré pour l'expérimentation en tal. 2006.

W. Wilbur and K. Sirotkin. The automatic identification of stop words. *Journal of Information Science*, 18 :45–55, 1992.

D. Wong, M. Dong, and D. Hu. Machine translation based on translation corresponding tree structure. *Tsinghua Science Technology*, 11 :25–31, 2006.

J. Yang, Y. Zhang, L. Li, and X. Li. Yedda : A lightweight collaborative text span annotation tool. *arXiv preprint arXiv :1711.03759*, 2017.

Y. Yang and W. Wilbur. Using corpus statistics to remove redundant words in text categorization. *JASIS*, 47 :357–369, 1996.

C. Zhang, H. Wang, Y. Liu, D. Wu, Y. Liao, and B. Wang. Automatic keyword extraction from documents using conditional random fields. 4 :1169–1180, 2008.

S. Zhang, Y. Wu, and J. Chang. Survey of image recognition algorithms. pages 542–548, 2020.

Z. Zhong and H. Ng. Word sense disambiguation improves information retrieval. *50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference*, 1 :273–282, 2012.

ANNEXE A

Liste des référentiels utilisés pour l'enrichissement

Référentiel	Éditeur	Source	Format
ADICAP	Association ADICAP	http ://www.oncopathologie.com/Cyberpat/squelettes/dmde_avis/Adicap_v5-03.pdf	PDF
CCAM	Assurance maladie	http ://www.ameli.fr/accueil-de-la-ccam/telechargement/index.php	DBase
CIM10	OMS / VIDAL	http ://apps.who.int/classifications/icd10/browse/2008/fr	Mysql
Cimo-3-topo	OMS	http ://apps.who.int/iris/bitstream/10665/43859/1/9789242545340_fre.pdf	Pdf
Cimo-3-morpho / Snomed-M	IHTSDO / ASIP Santé	https ://esante.gouv.fr/media/2605	Excel
VIDAL	VIDAL	http ://update.vidalhoptimal.fr/	Mysql
COSMIC	Cosmic	http ://cancer.sanger.ac.uk/cosmic	Excel
CTCAE	FFCD	http ://www.cepd.fr/CUSTOM/CEPD_toxicite.pdf	PDF
Protocole	Unicancer	Pas non disponible en ligne (fichier ad hoc fourni par Unicancer)	Excel

ANNEXE B

Exemples de comptes rendus médicaux

ECHOGRAPHIE DE LA FESSE GAUCHE **APPAREIL :** Echographe Siemens Antares - juin 2002 - N° Agrément : NL 00592.00.069

Nos références : *RD/MD*

INDICATION :
- Bilan d'extension métastatique.
- Patiente traitée pour cancer bronchique.

RESULTAT :
- On ne trouve pas la lésion focalisée ponctionnable.
- Les tissus mous de la fesse gauche sont homogènes, sans signe d'infiltration ou de masse tissulaire.

CONCLUSION :
On ne trouve aucune cible pour la ponction.
Il fallait consulter les images de l'IRM.

FIGURE B.1 – Exemple °1 d'un compte rendu.

21/02/2012 Consultation d'annonce Docteur M. CAMPITELLI
MLN

Fait le 21/02/2012 par le Docteur M. CAMPITELLI

Rappel : Octobre 2011 : biopsie de l'endomètre : adénocarcinome de type endométioïde sur métrorragies post-ménauposiques. HPV négatif. CA 125 à 53.
CHL, annexectomie bilatérale et curage iliaque externe bilatéral le 10.01.2012 :
Adénocarcinome infiltrant de type endométrioïde de l'endomètre de grade III.
Taille tumorale : 50 mm.
Infiltration de 80% du myomètre avec respect de la séreuse.
Présence d'images d'embolies tumorales.
Extension tumorale à l'isthme , l'endocol et au paramètre droit.
Les deux cornes , paramètre gauche, et l'exocol sont indemnes.
Les deux annexes droite et gauche sont peu modifiées.
- Curage iliaque externe droit :
Neuf ganglions indemnes.
- Curage iliaque externe gauche :
Quinze ganglions indemnes (15N-).
Cytopérinéale négative.

Patiente de 65 ans qui a été adressée pour prise en charge post-opératoire d'un adénocarcinome de l'endomètre.

Des métrorragies post-ménauposiques ont amenées à la découverte d'un adénocarcinome endométrioïde.

Bilan complémentaire :

IRM pelvienne du 04.11.2011 : lésion du col utérin avec perte de contours de l'anneau fibreux du col en antéro-latéral droite avec présence de ganglions iliaques externes de taille significative. Lésion mieux circonscrite de l'endomètre (strictement intra-endométriale)

Scanner TAP du 2.12.2011 normal.

TEP scanner du 6.12.2011 : Deux lésions modérément hypermétaboliques du fond et du col utérin.
Ganglions iliaques externes bilatéraux non hypermétaboliques.

Mammographie le 2.12.2011 classée ACR2

CHL, annexectomie bilatérale et curage iliaque externe bilatéral le 10.01.2012 :
Adénocarcinome infiltrant de type endométrioïde de l'endomètre de grade III.
Taille tumorale : 50 mm.
Infiltration de 80% du myomètre avec respect de la séreuse.
Présence d'images d'embolies tumorales.
Extension tumorale à l'isthme , l'endocol et au paramètre droit.
Les deux cornes , paramètre gauche, et l'exocol sont indemnes.
Les deux annexes droite et gauche sont peu modifiées.

FIGURE B.2 – Exemple °2 d'un compte rendu.

- Curage iliaque externe droit :
Neuf ganglions indemnes.
- Curage iliaque externe gauche :
Quinze ganglions indemnes (15N-).
Cytopérinéale négative.

Légère incontinence qui est précédent à la chirurgie.

Antécédents de varices des membres inférieurs avec port de bas de contention.

Antécédent de diabète.

Pas d'antécédents de phlébites ou de TVP

Signale une douleur de type neuropathique au niveau de la fesse gauche qui irradie de long de la jambe gauche et qui se calme au repos, gênant la marche. Apparue il y a 15 jours.

A l'examen : pas d'anomalie de la vulve ou de la marge anale en dehors d'un érythème modeste au niveau de la fourchette postérieure du vagin.

Vagin de longueur conservée, étroit. Cicatrice palpable du fond vaginal sans anomalie suspecte.

Examen au spéculum normal.

Décision :

Selon avis RCP, étant donnée la taille tumoral et l'index mitotique, indication à une chimiothérapie par CARBO TAXOL suivie d'une irradiation pelvienne limite supérieure L5-S1, limite inférieure au tiers inférieur du vagin. Dose pelvienne 45 Gy, 1,8 Gy par séance par tomothérapie. Suivie de curiethérapie vaginale à la dose de 15 Gy, du fait de l'atteinte du col.
Au total : Stade II, grade III avec embols N-
Patiente à revoir fin avril avec un scanner TAP, une prise de sang + CA 125

FIGURE B.3 – Exemple °3 d'un compte rendu.

21/02/2012 Consultation d'annonce Docteur M. CAMPITELLI
MLN

Fait le 21/02/2012 par le Docteur M. CAMPITELLI

Rappel : Octobre 2011 : biopsie de l'endomètre : adénocarcinome de type endométioïde sur métrorragies post-ménauposiques. HPV négatif. CA 125 à 53.
CHL, annexectomie bilatérale et curage iliaque externe bilatéral le 10.01.2012 :
Adénocarcinome infiltrant de type endométrioïde de l'endomètre de grade III.
Taille tumorale : 50 mm.
Infiltration de 80% du myomètre avec respect de la séreuse.
Présence d'images d'embolies tumorales.
Extension tumorale à l'isthme , l'endocol et au paramètre droit.
Les deux cornes , paramètre gauche, et l'exocol sont indemnes.
Les deux annexes droite et gauche sont peu modifiées.
- Curage iliaque externe droit :
Neuf ganglions indemnes.
- Curage iliaque externe gauche :
Quinze ganglions indemnes (15N-).
Cytopérinéale négative.

Patiente de 65 ans qui a été adressée pour prise en charge post-opératoire d'un adénocarcinome de l'endomètre.

Des métrorragies post-ménauposiques ont amenées à la découverte d'un adénocarcinome endométrioïde.

Bilan complémentaire :

IRM pelvienne du 04.11.2011 : lésion du col utérin avec perte de contours de l'anneau fibreux du col en antéro-latéral droite avec présence de ganglions iliaques externes de taille significative. Lésion mieux circonscrite de l'endomètre (strictement intra-endométriale)

Scanner TAP du 2.12.2011 normal.

TEP scanner du 6.12.2011 : Deux lésions modérément hypermétaboliques du fond et du col utérin.
Ganglions iliaques externes bilatéraux non hypermétaboliques.

Mammographie le 2.12.2011 classée ACR2

CHL, annexectomie bilatérale et curage iliaque externe bilatéral le 10.01.2012 :
Adénocarcinome infiltrant de type endométrioïde de l'endomètre de grade III.
Taille tumorale : 50 mm.
Infiltration de 80% du myomètre avec respect de la séreuse.

FIGURE B.4 – Exemple °4 d'un compte rendu.

21/11/2012 Scanner thorax abdomen pelvis

Docteur S. NEUENSCHWANDER

/ PMN

Examen : Scanner I.P.C. Scanner thorax et abdopelvien

Indications :

Bilan confrontation avec le scanner de juillet 2012. Lésion de l'endometre en octobre 2011. Hysterectomie totale et curage ganglionnaire et radiothérapie pelvienne.

Technique :

On réalise une série sur le foie sans injection suivie d'une grande hélice veineuse thoraco abdomino pelvienne.
Quantité : (XENETIX 350 91 ml (lot n°12wf018a) Débit de 3ml par seconde.
Scanner SIEMENS Somaton AS20 ; Agrément 75/056/028/M/01/2010 Scanner thoraco-abdominopelvien = 2 099,00 mGy x cm

Résultats :

Abdomen :
Le foie est de taille normale. Sa densité est homogène avec un micro kyste du 8 de 5 mm.
Les vaisseaux sont perméables. Les voies biliaires sont fines. Micro lithiase vésiculaire.
Pas d'adénopathie dans le pédicule hépatique et la région coeliaque.
Pancréas de taille normale. Sa densité est homogène.
Rate normale.
Surrénales Reins et Rétro-péritoine : Les reins sont de taille normale avec des cavités fines.
Kyste des deux reins.
Les deux surrénales sont normales. Pas d'adénopathie dans le rétropéritoine.
Pelvis :
Pas d'adénopathie le long des deux axes iliaques. Lymphocele gauche banale de10,8 cm. Pas d'ascite.

Thorax :
Médiastin : pas d'adénopathie, pas d 'anomalie pleurale.
Parenchyme : micro calcification banale de la base droite.

Os :
Pas d'anomalie individualisée sur les coupes réalisées.

CONCLUSION :

Scanner thoraco-abdominal et pelvien pouvant être considéré comme normal.

Docteur Odile GHEMARD

FIGURE B.5 – Exemple °5 d'un compte rendu.

29/11/2011	Consultation de surveillance en radiothérapie du sein	Docteur M. CAMPITELLI

MLN

Fait le 22/10/2010 par le Docteur M. CAMPITELLI

Rappel : 53 ans; Découverte opacité spiculée QSE sein G 5 mm. Biopsies +
28/04/2008 Mastectomie partielle G avec curage axillaire, Examen extemporané
Carcinome canalaire infiltrant de 19 mm - grade II - index mitotique faible; Embols-;
Marges saines; 7 N-
RO+; RP+, Her2 -
8/8 au 8/10/2008 radiothérapie mammaire gauche 50 Gy + boost 20 Gy QSE (exérèse étroite berge distale)
10/2008 TAMOXIFENE
10/2010 FEMARA

Examen clinique : La palpation mammaire est normale.
Les aires ganglionnaires sont libres.

3 ans de recul.

Sous FEMARA, avec bonne tolérance.

Résultat esthétique (seulement si traitement conservateur) :	Bon
Evaluation :	Oui

Décision :

RV en avril 2012 avec Docteur LEVIEIL avec mammographie et échographie mammaire (ordonnance remise à la patiente).
Sera revue par le Docteur CAMPITELLI dans un an.
Prescription d'une ostéodensitométrie pour dans 6 mois.

Information donnée : Compte-rendu dicté en présence de la patiente

FIGURE B.6 – Exemple °6 d'un compte rendu.

Docteur Jean CHERBIT 35 B BD EMILE COMBES
13200 ARLES

Marseille, le 09/10/2012

N° dossier : null0359741129013604775658

Cher Confrère,

Je vous prie de trouver ci-joint le compte-rendu de consultation de votre patiente Mme TOVMASYAN née LAYTON JOVAN âgée de 49 ans.

En cas d'urgence, vous pouvez contacter de 8 h 30 à 18 h 30 un
Oncologue de l'Institut au numéro suivant : 06.61.02.68.87.

Bien confraternellement.

Docteur ANNE MADROSZYK

Ces documents ont été relus et validés électroniquement par le médecin signataire.

NOM USUEL : TOVMASYAN NOM DE JEUNE FILLE : LAYTON
PRENOM :JOVAN SEXE : F
DATE DE NAISSANCE : 21/03/1963 N° DOSSIER : null0359741129013604775658
DATE DE CONSULTATION : 09/10/2012
Réf. AM/ CAT BOU

Motif de Consultation
Patiente revue dans le cadre de son traitement par HERCEPTIN associé à l?ARIMIDEX pour un adénocarcinome mammaire avec antécédent de lésions secondaires hépatiques actuellement en rémission.
Observations
Elle devait initialement réaliser un scanner cérébral et thoraco-abdominal mais a totalement omis de le réaliser.
Par ailleurs échographie- mammographie qui a pu être décalée, réalisée ce jour qui est tout à fait satisfaisante.
La patiente est dans un excellent état général, à noter qu?elle n?a pas non plus réalisé son échographie cardiaque.

Décision Thérapeutique
Réalisation ce jour de son injection d?HERCEPTIN 6 mg/kg toutes les 3 semaines.
Renouvellement de l?ARIMIDEX.
Programmation d?un nouveau scanner cérébral et thoraco-abdominal dès que possible.
La patiente doit reprendre contact auprès de son cardiologue pour l?échographie cardiaque.
Poursuite des injections toutes les 3 semaines en attendant.

Courrier dicté en présence du patient.

FIGURE B.7 – Exemple °7 d'un compte rendu.

Marseille, le 12/09/2013

N° dossier : nullfa5a64ea3f357a6f24e765041aba58172ea1b4cf2acc6ed6d0fa2a71458d890f

Cher Confrère,

Je vous prie de trouver ci-joint le compte-rendu de consultation de votre patient M. ALECCA TAKISHA âgé de 64 ans.

Bien confraternellement.

Docteur M Jean-François MOULIN

Ces documents ont été relus et validés électroniquement par le médecin signataire.

NOM USUEL : ALECCA NOM DE JEUNE FILLE :
PRENOM :TAKISHA SEXE : H
DATE DE NAISSANCE : 28/10/1948 N° DOSSIER :
nullfa5a64ea3f357a6f24e765041aba58172ea1b4cf2acc6ed6d0fa2a71458d890f
DATE DE CONSULTATION : 12/09/2013
Réf. JFM/NM 2

Motif de Consultation
Prise en charge diagnostique et thérapeutique.

Antécédents médicaux : glaucome.

Antécédents chirurgicaux: néant.

Antécédents familiaux :
Cancer du poumon chez son père,
Cancer du sein chez sa mère.

Toxiques : Tabac : 70 PA arrêté il y a 3 semaines,
Alcool : 1 verre de vin par repas.

Social : Marié, trois filles à proximité,
Retraité, ancien gendarme puis commercial.

Traitement en cours : Néant.

Histoire de la maladie :
Le patient est hospitalisé en urgence pour dyspnée et douleurs thoraciques, en rapport avec un épanchement pleural droit. Une fibroscopie bronchique est réalisée dont je n?ai pas les résultats.
La ponction pleurale, réalisée le 29/08/2013, montre un liquide hémorragique inflammatoire sans cellule suspecte de malignité.
Un scanner TAP, dont je n?ai pas le compte rendu, retrouve un épanchement pleural droit, une tumeur vésicale du trigone à droite, sans dilation des cavités pyélocalicielles et une adénopathie latéro aortique gauche.
La cystoscopie retrouve 2 lésions endovésicales, l?une superficielle médiane de la face postérieure, réséquée en totalité, la seconde, plus volumineuse d?allure infiltrante, centrée sur l?orifice urétéral droit, également réséquée en totalité jusqu?au muscle.

FIGURE B.8 – Exemple °8 d'un compte rendu.

Marseille, le 12/09/2013

N° dossier : nullfa5a64ea3f357a6f24e765041aba58172ea1b4cf2acc6ed6d0fa2a71458d890f

Cher Confrère,

Je vous prie de trouver ci-joint le compte-rendu de consultation de votre patient M. ALECCA TAKISHA âgé de 64 ans.

Bien confraternellement.

Docteur M Jean-François MOULIN

Ces documents ont été relus et validés électroniquement par le médecin signataire.

NOM USUEL : ALECCA NOM DE JEUNE FILLE :
PRENOM :TAKISHA SEXE : H
DATE DE NAISSANCE : 28/10/1948 N° DOSSIER :
nullfa5a64ea3f357a6f24e765041aba58172ea1b4cf2acc6ed6d0fa2a71458d890f
DATE DE CONSULTATION : 12/09/2013
Réf. JFM/NM 2

Motif de Consultation
Prise en charge diagnostique et thérapeutique.

Antécédents médicaux : glaucome.

Antécédents chirurgicaux: néant.

Antécédents familiaux :
Cancer du poumon chez son père,
Cancer du sein chez sa mère.

Toxiques : Tabac : 70 PA arrêté il y a 3 semaines,
Alcool : 1 verre de vin par repas.

Social : Marié, trois filles à proximité,
Retraité, ancien gendarme puis commercial.

Traitement en cours : Néant.

Histoire de la maladie :
Le patient est hospitalisé en urgence pour dyspnée et douleurs thoraciques, en rapport avec un épanchement pleural droit. Une fibroscopie bronchique est réalisée dont je n?ai pas les résultats.
La ponction pleurale, réalisée le 29/08/2013, montre un liquide hémorragique inflammatoire sans cellule suspecte de malignité.
Un scanner TAP, dont je n?ai pas le compte rendu, retrouve un épanchement pleural droit, une tumeur vésicale du trigone à droite, sans dilation des cavités pyélocalicielles et une adénopathie latéro aortique gauche.
La cystoscopie retrouve 2 lésions endovésicales, l?une superficielle médiane de la face postérieure, réséquée en totalité, la seconde, plus volumineuse d?allure infiltrante, centrée sur l?orifice urétéral droit, également réséquée en totalité jusqu?au muscle.

FIGURE B.9 – Exemple °9 d'un compte rendu.

21/02/2012 Consultation d'annonce Docteur M. CAMPITELLI
MLN

Fait le 21/02/2012 par le Docteur M. CAMPITELLI

Rappel : Octobre 2011 : biopsie de l'endomètre : adénocarcinome de type endométioïde sur métrorragies post-ménauposiques. HPV négatif. CA 125 à 53.
CHL, annexectomie bilatérale et curage iliaque externe bilatéral le 10.01.2012 :
Adénocarcinome infiltrant de type endométrioïde de l'endomètre de grade III.
Taille tumorale : 50 mm.
Infiltration de 80% du myomètre avec respect de la séreuse.
Présence d'images d'embolies tumorales.
Extension tumorale à l'isthme , l'endocol et au paramètre droit.
Les deux cornes , paramètre gauche, et l'exocol sont indemnes.
Les deux annexes droite et gauche sont peu modifiées.
- Curage iliaque externe droit :
Neuf ganglions indemnes.
- Curage iliaque externe gauche :
Quinze ganglions indemnes (15N-).
Cytopérinéale négative.

Patiente de 65 ans qui a été adressée pour prise en charge post-opératoire d'un adénocarcinome de l'endomètre.

Des métrorragies post-ménauposiques ont amenées à la découverte d'un adénocarcinome endométrioïde.

Bilan complémentaire :

IRM pelvienne du 04.11.2011 : lésion du col utérin avec perte de contours de l'anneau fibreux du col en antéro-latéral droite avec présence de ganglions iliaques externes de taille significative. Lésion mieux circonscrite de l'endomètre (strictement intra-endométriale)

Scanner TAP du 2.12.2011 normal.

TEP scanner du 6.12.2011 : Deux lésions modérément hypermétaboliques du fond et du col utérin.
Ganglions iliaques externes bilatéraux non hypermétaboliques.

Mammographie le 2.12.2011 classée ACR2

FIGURE B.10 – Exemple °10 d'un compte rendu.

CRLC Val d'Aurelle

Résumé de soins infirmiers le 30/09/2007

PATIENT : null13534
NETHERCUTT LOREEN né(e) le : 26/11/1953

Service : Chirurgie A2 Tél : 04-67-61-31-05 Fax : 04-67-61-37-03
Secrétariat Soins Externes : 04-67-61-30-38 IDE Soins Externes : 04-67-61-37-30
Médecin référent : Bernard SAINT-AUBERT

Contacts extérieurs :

Opéré(e) le 26/09/2007 pour : Curage axillaire gauche et parage de plaie superficielle < 3 cm.

Cibles Données Actions Résultats

Soins de base

Soins techniques

pansement cycle bétadine et cicaplaie.

Soins relationnels et éducatifs

BMR
Isolement : NON

Examens :

Rendez-vous :

Commentaire : Réfection du pansement ce jour avec ablation du drain axillaire gauche. Papiers de sortie et informations donnés ainsi que le traitement pour la journée. Dossier radiologique rendu.

Catherine Remp LUCIANI

IDE

FIGURE B.11 – Exemple °11 d'un compte rendu.

BG /GL
N° DOSSIER : null29441

Cher Confrère,

Monsieur RONNIE ORK, né(e) le 20/03/1939 (74 ans),a été hospitalisé(e) en Hôpital de jour le 11/07/2013 pour le j8 de chimiothérapie par TAXOL CARBOPLATINE.

Meilleur état général, mais recrudescence des douleurs de la hanche.
Amélioration du taux d'hémoglobine qui est à 10.2 gr/dl ce jour.

Ce jour, l'état général est OMS 2.

Après entretien avec le patient et examen clinique, en l'absence de contre-indication clinique et biologique, délivrance du traitement suivant :
TAXOL 80
CARBOPLATINE AUC2.
La tolérance immédiate du traitement a été bonne.

Traitement de sortie :
ZOPHREN
Remise d'une ordonnance de IOMERON

Monsieur ORK reviendra le 25.07.13 en Hôpital de jour pour le j1 de la cure 2.

Bien cordialement.

Docteur Blandine GALLET-SUCHET

Courrier expédié à :

Docteur DOMINIQUE SARRABERE
- 22, route de Lavérune - 34070 MONTPELLIER

FIGURE B.12 – Exemple °12 d'un compte rendu.

Montpellier, le 12/10/2006

GENRE le TITREL PRENOM NOM ADR1 ADR2 ADR3 ADR4
CP VILLE

MY /VL null9641

N° réseau ONCOLR :

Cher Confrère,

Je vous prie de bien vouloir trouver, ci-joint, le compte-rendu d'hospitalisation en Hôpital de jour de

Monsieur JANE DREWRY, né(e) le 23/08/1938 (68 ans),

du 02/10/2006

Je vous prie de croire, Cher Confrère, en l'expression de mes sentiments cordiaux et dévoués.

Professeur Marc YCHOU

Courrier à l'attention du :
Docteur Hélène FANTON Cabinet de Médecine Générale 15 Route de Lavérune 34990 JUVIGNAC

FIGURE B.13 – Exemple °13 d'un compte rendu.

ANNEXE C

Règles pour la détection des dates

```
1
2     # Lookbehind négatif pour les textes antérieurs à une date
3     DATE_NEG_LOOKBEHIND = r"(?=(?<!\bnée ))(?=(?<!\bné ))(?=(?<!\bnée le ))
      (?=(?<!\bné le ))(?=(?<!\bnee ))(?=(?<!\bne ))(?=(?<!\bnee le ))(?=(?<!\
      bne le ))(?=(?<!\bnee en ))(?=(?<!\bne en ))(?=(?<!\bnee en ))(?=(?<!\
      bne en ))"
4
5     # Match les années seules entre 1950-2059 précédées par {en; depuis; de
      }
6     REGEX_YEAR = r"(((\b(?<!\bne)(?<!\bnee)(?<!\bné)(?<!\bnée)\b
      |(?<=[,;\(-]))\s+en|(?<=depuis)|(?<=annee)|(?<=année)|(?<=\bet)|(?<=dé
      but)|(?<=debut)|(?<=fin))|^( |-|\*|\.)*\s*en)\s
      +(19[5-9][0-9]|20[0-5][0-9])\b"
7
8     # Match les années seules au début des phrases (2014: métastase
      pulmonaire)
9     REGEX_YEAR_BEGINNING = r"^( |-|\*|\.)*\b(19[5-9][0-9]|20[0-5][0-9])\b\s
      *:?"
10
11    # Faire correspondre les séquences mois/année avec le mois écrit en
      chiffres. .
12    REGEX_MONTH_DIGIT_YEAR = (date_neg_lookbehind + r"(" + ''.join([r"(\b
      (0?[1-9]|1[0-2])\b\s*" + x + r"?\s*\b(((19)?[5-9][0-9])|((20)
      ?[0-5][0-9]))\b)|" for x in regex_date_separators]))[:-1] + r")"
13
14    # Faire correspondre les séquences mois / année avec le mois écrit en
      lettres
15    REGEX_MONTH_LETTER_YEAR = date_neg_lookbehind + r"\b((janv(\.|\b|ier\b)
      )|(f[ée]v(r)?(\.|\b|ier\b))|(mar(\.|\b|s\b))|(avr(\.|\b|il\b))|(mai)\b|(
      jui(\.|\b|n\b))|(juil(\.|\b|let\b))|(ao[uû](\.|\b|t\b))|(sept(\.|\b|
      embre\b))|(oct(\.|\b|obre\b))|(nov(\.|\b|embre\b))|(d[ée]c(\.|\b|embre\b
      )))\s*\b(((19)?[5-9][0-9])|((20)?[0-5][0-9]))"
16
17    # Date complète avec mois écrit en chiffres
18    REGEX_DAY_MONTH_DIGIT_YEAR = (date_neg_lookbehind + ''.join([r"(\b
      (0?[1-9]|[1-2][0-9]|3[0-1])(\s*" + x + r"\s*)(0?[1-9]|1[0-2])(\s*" + x +
       r"\s*)((19)?[5-9][0-9]|(20)?[0-5][0-9])\b)|" for x in
      regex_date_separators]))[:-1]
19
```

```
    # Date complète avec mois écrit en lettres
    REGEX_DAY_MONTH_LETTER_YEAR = (date_neg_lookbehind + ''.join([r"(\b
    (0?[1-9]|[1-2][0-9]|3[0-1])(\s*" + x + r"\s*)\b((janv(\.|\b|ier\b))|(f[é
    e]v(r)(\.|\b|ier\b))|(mar(\.|\b|s\b))|(avr(\.|\b|il\b))|(mai)\b|(jui
    (\.|\b|n\b))|(juil(\.|\b|let\b))|(ao[uû](\.|\b|t\b))|(sept(\.|\b|embre\b
    ))|(oct(\.|\b|obre\b))|(nov(\.|\b|embre\b))|(d[ée]c(\.|\b|embre\b)))(\s*
    " + x + r"\s*)((19)?[5-9][0-9]|(20)?[0-5][0-9])\b)|" for x in
    regex_date_separators]))[:-1]


    # Faire correspondre les séquences jour/mois avec le mois écrit en
    lettres
    REGEX_DAY_MONTH_LETTER = (date_neg_lookbehind + ''.join([r"(\b
    (0?[1-9]|[1-2][0-9]|3[0-1])\b(\s*" + x + r"?\s*)\b((janv(\.|\b|ier\b))|(
    f[ée]v(r)?(\.|\b|ier\b))|(mar(\.|\b|s\b))|(avr(\.|\b|il\b))|(mai)\b|(jui
    (\.|\b|n\b))|(juil(\.|\b|let\b))|(ao[uû](\.|\b|t\b))|(sept(\.|\b|embre\b
    ))|(oct(\.|\b|obre\b))|(nov(\.|\b|embre\b))|(d[ée]c(\.|\b|embre\b))))|"
    for x in regex_date_separators]))[:-1]

    # Faire correspondre les séquences jour/mois avec le mois écrit en
    chiffres
    REGEX_DAY_MONTH_DIGIT = (date_neg_lookbehind + date_pos_lookbehind + r"
    (" + ''.join([r"(\b(0?[1-9]|[1-2][0-9]|3[0-1])\b(\s*" + x + r"?\s*)\b
    (0?[15-9]|0[2-4]|1[0-2])\b)|" for x in regex_date_separators]))[:-1] + r
    ")"

    # Match mois seul avec indication
    REGEX_MONTH_ALONE = date_neg_lookbehind + r"\b(fin|d[eé]but) \b(janvier
    |f[ée]vrier|mars|avril|mai|juin|juillet|ao[uû]t|septembre|octobre|
    novembre|d[ée]cembre\b)"
```

Code C.1 – Expressions régulières pour la détection des dates

ANNEXE D

Liste des neuf protocoles ambigus

Protocole	**Description**
AC	Adriamycine, Cyclophosphamide
ACE	Adriamycine, Cyclophosphamide, Étoposide
BEP	Bléomycine, Étoposide ou VP-16, cisPlatine
CAP	Cyclophosphamide, Adriblastine, cisPlatine
CMF	Cyclophosphamide, méthotrexate, Fluoro-Uracile
EP	Étoposide + cisPlatine
FAC	Fluoro-Uracile, Adriamycine, cyclophosphamide
FUN	Fluoro-Uracile, Navelbine
VICE	Vincristine, Ifosfamide, Carboplatine, Étoposide