# TALMed : Traitement Automatique de la Langue Médicale

Adrien Bazoge

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

# THÈSE DE DOCTORAT DE

NANTES UNIVERSITÉ

ÉCOLE DOCTORALE N° 641
*Mathématiques et Sciences et Technologies du numérique de l'Information et de la Communication*
Spécialité : *Informatique*

Par
**Adrien BAZOGE**

# TALMed : Traitement Automatique de la Langue Médicale

**Thèse présentée et soutenue à Nantes, le 16 janvier 2024**
**Unité de recherche : UMR6004 – Laboratoire des Sciences et du Numérique de Nantes (LS2N)**

## Rapporteurs avant soutenance :

Laure SOULIER Maîtresse de conférences, Sorbonne Université
Didier SCHWAB Professeur des Universités, Université Grenobles Alpes

## Composition du Jury :

| | | |
|---|---|---|
| Président : | Gayo DIALLO | Professeur des Universités, Université de Bordeaux |
| Examinateurs : | Laure SOULIER | Maîtresse de conférences, Sorbonne Université |
| | Didier SCHWAB | Professeur des Universités, Université Grenobles Alpes |
| | Gayo DIALLO | Professeur des Universités, Université de Bordeaux |
| Dir. de thèse : | Emmanuel MORIN | Professeur des Universités, Nantes Université |
| Co-dir. de thèse : | Béatrice DAILLE | Professeur des Universités, Nantes Université |
| | Pierre-Antoine GOURRAUD | Professeur des Universités-Praticien Hospitalier, Nantes Université |

# Remerciements

Je tiens tout d'abord à remercier Emmanuel Morin, Béatrice Daille et Pierre-Antoine Gourraud, pour leur confiance depuis mon alternance en M2 et pour m'avoir proposé de poursuivre en thèse sur ce sujet passionnant. Je remercie Emmanuel Morin pour avoir dirigé ma thèse et pour son soutien sans faille tout au long de ce périple tumultueux qu'est une thèse. Merci d'avoir fait en sorte que ces trois années de thèse soient les plus agréables possibles. Je remercie Béatrice pour avoir co-dirigé ma thèse et pour son aide précieuse lors de la rédaction de ce manuscrit. Je remercie Pierre-Antoine pour avoir co-encadré ma thèse, pour son dynamisme enthousiasmant et ses précieux conseils qui m'ont aidé à prendre du recul sur mon travail.

Je remercie les membres du jury, Laure Soulier et Didier Schwab, d'avoir accepté d'être rapporteurs de ma thèse, ainsi que Gayo Diallo de faire partie de mon jury en tant qu'examinateur.

Je remercie également les membres de mon CSI, Harold Mouchère et Pierre Zweigenbaum, de m'avoir suivi et conseillé pendant ces trois années.

Je remercie les membres de l'équipe TALN du LS2N avec qui j'ai partagé le quotidien, et plus particulièrement les doctorants pour ces moments d'échanges (scientifiques ou non) et de procrastination.

Je remercie toute l'équipe de la Clinique des données du CHU de Nantes pour m'avoir introduit au domaine médical et m'avoir accompagné durant ces quatre dernières années.

Je remercie également toutes les personnes avec qui j'ai pu collaborer durant ces trois années, sans qui les travaux présentés dans ce manuscrit n'auraient pu voir le jour. Merci à Matthieu Wargny, Samy Hadjadj et aux nombreux annotateurs pour le projet GAVROCHE, ainsi qu'à Julien Herbert pour son aide dans le déploiement du projet à l'interrégion. Merci à Pacôme Constant dit Beaufils et à Matilde Karakachoff pour le projet sur les déterminants sociaux de santé. Et enfin, merci à Yanis Labrak, Richard Dufour et Mickael Rouvier pour nos nombreuses collaborations qui m'ont beaucoup apporté.

Enfin, je remercie ma famille et mes amis (Amine, Lucas et Kévin) pour leur soutien.

# TABLE DES MATIÈRES

# INTRODUCTION

## 0.1 Contexte

Avec l'informatisation des données de santé, les établissements hospitaliers ont été amenés à changer de support de stockage des informations, passant de dossiers papier à des dossiers médicaux électroniques ou dossiers patients informatisés (DPI) (Gunter and Terry, 2005; Adler-Milstein et al., 2017). Ces dossiers médicaux sont enrichis de manière longitudinale, lors de la prise en charge des patients, avec des données provenant de différentes sources et applications métiers, telles que les prescriptions médicamenteuses, l'imagerie, les examens biologiques, les hospitalisations et les consultations. Toutes ces informations sont centralisées et constituent le système d'information hospitalier (SIH). Initialement, l'archivage de ces données avait pour usage primaire le suivi de l'état de santé des patients et leur prise en charge dans le cadre du soin (Safran et al., 2007). Cependant, la variété et le volume des données collectées au fil des années ont permis l'émergence d'usages secondaires, qui ne concernent pas directement la prise en charge des patients, tels que la recherche, le pilotage des établissements, l'évaluation de la qualité des soins et la veille sanitaire (Meystre et al., 2017). L'intérêt porté à ces usages secondaires a été grandement accentué ces dernières années avec la crise de la Covid-19 (Chen et al., 2020a). Néanmoins, pour une réutilisation systématique des données de santé à large échelle, notamment dans la recherche avec les études cliniques et épidémiologiques, il est nécessaire que ces données soient harmonisées et standardisées (Haarbrandt et al., 2016; Gagalova et al., 2020b). Or les données stockées dans les SIH sont très hétérogènes car basées sur des formats et des terminologies non standardisés issus de différentes solutions logicielles métiers.

La mise en place des entrepôts de données de santé (EDS) a permis d'établir un lien entre les données collectées dans le cadre du soin et leur réutilisation dans les usages secondaires (Safran et al., 2007). Bien que leur mise en place ait été motivée en premier lieu par les usages secondaires proches du soin et du pilotage des établissements, c'est aujourd'hui la réutilisation des données à des fins de recherche qui incitent les établissements à se doter de ces entrepôts (Commission nationale informatique et des libertés, 2021). Les EDS intègrent les données issues des différents logiciels métiers sources et les standardisent

dans un schéma commun avec éventuellement des nomenclatures communes. On retrouve dans les EDS des données similaires à celles stockées dans les SIH : principalement les données structurées (données administratives, diagnostics et procédures codées, biologie, etc.), les données textuelles et plus rarement les données d'imagerie ou de génomique, car plus complexes et volumineuses. Bien que ces EDS soient devenus les points d'entrée pour l'usage des données de santé pour la recherche, accéder à ces ressources doit être justifié par un projet de recherche bien défini et seules les données strictement nécessaires au projet sont mises à disposition.

Des alternatives aux EDS ont été proposées pour faciliter l'accès aux données de santé pour la recherche. C'est le cas de la base de données *Multiparameter Intelligent Monitoring in Intensive Care* (MIMIC) (Moody and Mark, 1996) dont l'objectif est de constituer une collection de données désidentifiées de patients admis en soins intensifs au Centre médical des diaconesses Beth Israel à Boston, États-Unis. Avec une première version publiée en 1996, MIMIC continue d'être développé et amélioré au fil du temps et capture un ensemble très large de données collectées dans le cadre du soin des patients tels que des comptes rendus cliniques et des images radiologiques. Les versions les plus récentes et couramment utilisées, MIMIC-III (Johnson et al., 2016) publié en 2015 et MIMIC-IV (Johnson et al., 2020) en 2020, rassemblent plusieurs dizaines de milliers de patients admis entre 2001 et 2012 pour MIMIC-III et entre 2008 et 2019 pour MIMIC-IV. Ces bases de données sont accessibles gratuitement à tous les chercheurs dans le monde, peu importe leur domaine de recherche, ce qui en font des ressources précieuses car elles offrent des informations très détaillées sur une grande population de patients telles que les signes vitaux, les résultats de laboratoires et les médicaments. Toutes ces données permettent de réaliser différentes études analytiques en épidémiologie, d'améliorer les décisions cliniques et de développer des outils et logiciels pour mieux traiter ces données. D'autres bases de données ont été rendus accessibles suivant le modèle proposé par MIMIC, on peut citer : *eICU Collaborative Research Database* (eICU-CRD) (Pollard et al., 2018), *Amsterdam University Medical Center data base* (AmsterdamUMCdb) (Thoral et al., 2021) et *High time-resolution intensive care unit dataset* (HiRID) (Faltys et al., 2021).

La majorité des informations cliniques sont collectées au travers de comptes rendus textuels et sont stockées dans les EDS sous forme de textes libres. Ce format de stockage des informations amène à développer des solutions s'appuyant sur le traitement automatique des langues pour exploiter au mieux ces données, notamment pour en extraire des informations. Ces dernières années, les modèles de langue pré-entraînés permettent d'ob-

tenir d'excellents résultats dans la plupart des tâches du TAL. La majorité des modèles sont pré-entraînés sur des corpus du domaine général, tels que des livres, des encyclopédies ou des articles journalistiques. Bien que ces modèles puissent être utilisés dans divers contextes, il est nécessaire de les adapter pour obtenir des performances optimales dans les domaines spécialisés, tels que la finance (Yang et al., 2020b), le tourisme (Zhu et al., 2021) ou le médical (Yang et al., 2022). L'adaptation d'un modèle à un domaine de spécialité n'est pas aisé, car limité par la quantité de ressources disponibles dans la langue et le domaine cible. Pour le domaine biomédical, des solutions ont été proposées dans la langue anglaise pour adapter ces modèles grâce à des corpus spécialisés suffisamment fournis, tels que les comptes rendus textuels de la base de données MIMIC (Huang et al., 2020) ou les résumés d'articles scientifiques issus de PubMed (Gu et al., 2021). Dans les langues moins dotées que l'anglais, les modèles de langue pré-entraînés pour le domaine biomédical se font plus rares, dû au manque de corpus pour entraîner et évaluer ces modèles.

## 0.2 Contributions

C'est dans ce contexte que s'inscrit ce travail de thèse. Nous proposons, dans un premier temps, d'adapter au domaine médical divers modèles de langue pré-entraînés pour la langue française. Nous nous intéressons tout d'abord aux modèles basés sur l'architecture BERT (Devlin et al., 2019), qui se sont imposés ces dernières années comme l'état de l'art. Pour adapter au mieux ce type de modèle au domaine médical, nous présentons une étude évaluant, sur un ensemble de tâches biomédicales, l'impact des données utilisées pour l'apprentissage, incluant des données cliniques issues d'établissements hospitaliers et des données médicales publiques récoltées sur le web. Nous étudions également l'impact de la stratégie d'apprentissage, en comparant les modèles entraînés à partir de zéro aux modèles entraînés à partir d'un modèle général existant. El Boukkouri et al. (2022) ont évalué l'impact de ces deux stratégies pour les modèles anglais dans le domaine biomédical et ont montré qu'il est préférable de continuer le pré-entraînement d'un modèle existant. Notre objectif est de vérifier si nous arrivons aux mêmes conclusions pour les modèles français. Nous nous intéressons ensuite aux modèles basés sur l'architecture Longformer (Beltagy et al., 2020), qui permettent de travailler sur des textes longs. En effet, les documents cliniques sont souvent très longs, avec parfois plusieurs milliers de mots, et les modèles basés sur l'architecture BERT ne permettent pas d'exploiter la totalité d'un document

puisqu'ils sont limités à 512 *tokens*[1] en entrée. De la même manière que pour l'adaptation des modèles BERT, nous entraînons un modèle Longformer sur des données récoltées sur le web et évaluons ce modèle sur des corpus biomédicaux composés de documents longs.

L'évaluation extrinsèque est aussi un aspect important dans la construction de modèles de langue et nécessite d'avoir des jeux de données sur des tâches variées pour bien mesurer la qualité des modèles pré-entraînés. Cependant, les jeux de données disponibles en français pour le domaine biomédical sont très limités et peu variés dans les tâches qu'ils proposent. L'accessibilité de jeux de données pour la recherche est un point essentiel, à la fois pour la reproductibilité des expériences, mais aussi pour le développement de nouveaux outils de traitement automatique. Pour faciliter l'évaluation des modèles de langue et permettre la reproductibilité des expériences au sein de la communauté, nous introduisons DrBenchmark, un benchmark reprenant les corpus biomédicaux existants dans la littérature scientifique. Deux nouveaux corpus sont aussi introduits dans ce benchmark : FrenchMedMCQA (Labrak et al., 2022), un corpus de questions à choix multiples (QCM) issues de concours d'internat en pharmacie et DiaMed, un corpus de cas cliniques annoté avec des codes CIM-10 (Classification internationale des maladies).

L'usage d'outils de TAL en contexte clinique, tels que la recherche ou le soin, doit être fait avec précaution. En effet, les systèmes ne sont jamais parfaits, y compris les derniers systèmes état de l'art, et peuvent extraire ou déduire des informations incorrectes. De plus, les systèmes de TAL sont généralement évalués sur des annotations manuelles, alors que d'un point de vue clinique, les études travaillent au niveau du patient ou du séjour (Velupillai et al., 2018). Afin de mieux aborder la réutilisation des modèles pré-entraînés dans un cadre d'application clinique réel, nous explorons l'usage de ces modèles en contexte de recherche clinique à travers de deux cas d'usages d'extraction d'informations. Le premier cas d'usage est le projet GAVROCHE, une étude multicentrique qui s'intéresse à l'association entre la variabilité glycémique des premiers jours suivant l'admission et le décès chez des patients hospitalisés pour insuffisance cardiaque aiguë. Dans ce projet, l'objectif est d'extraire automatiquement, à partir des comptes rendus d'hospitalisation, une série de variables prédéfinies pour phénotyper les patients de l'étude. Pour cela, nous présentons un corpus annoté sur lequel nous appliquons les modèles de langue introduits dans cette thèse. Le deuxième cas d'usage porte sur l'extraction des déterminants sociaux de santé dans les comptes rendus hospitaliers. Nous présentons un corpus annoté avec 13 déterminants sociaux sur lequel nous évaluons les modèles de langue introduits dans cette

1. un *token* est un sous-mot

thèse.

## 0.3 Plan du manuscrit

La suite de ce manuscrit est organisée en cinq chapitres. Dans le premier chapitre, nous présentons une revue systématique de la littérature sur l'usage du TAL dans les entrepôts de données de santé. Notre objectif est d'identifier les tâches de TAL utilisées pour traiter les documents textuels issus des EDS ainsi que les méthodes employées pour traiter ces tâches.

Dans le deuxième chapitre, nous étudions les différents types de représentations sémantiques des mots, allant des sacs de mots aux représentations contextualisées basées sur l'architecture Transformer (Vaswani et al., 2017). Nous décrivons ensuite les modèles de langue, en présentant les diverses architectures, les tâches permettant de les entraîner, ainsi que les modèles pré-entraînés existants et comment les adapter à un domaine de spécialité, en mettant l'accent sur les travaux portant sur le domaine médical. Nous faisons également un tour d'horizon des tâches utilisées pour l'évaluation extrinsèque des modèles en domaine biomédical et clinique, en illustrant avec des corpus pour chacune des tâches.

Dans le troisième chapitre, nous décrivons les corpus cliniques construits durant la thèse. La première section de ce chapitre est consacrée au corpus constitué dans le cadre du projet multicentrique GAVROCHE. Nous présentons l'objectif de ce corpus, le schéma d'annotation, ainsi que le processus d'annotation. Dans la deuxième section, nous présentons un corpus portant sur les déterminants sociaux de santé, composé de paragraphes issus de comptes rendus hospitaliers du CHU de Nantes. Nous décrivons le schéma et le processus d'annotation, ainsi que les résultats obtenus.

Le quatrième chapitre est consacré à l'adaptation des modèles de langue BERT (Devlin et al., 2019) et Longformer (Beltagy et al., 2020) aux domaines biomédical et clinique français. Dans ce chapitre, nous introduisons tout d'abord DrBenchmark, un benchmark de tâches biomédicales françaises permettant d'évaluer les modèles de langue pré-entraînés. Ce benchmark rassemble la majorité des corpus biomédicaux français publiés dans la littérature scientifique. Nous présentons ensuite DrBERT et ChuBERT, deux modèles de langues pré-entraînés, le premier sur des données biomédicales récoltées sur le web et le second sur des phrases simples extraites de comptes rendus hospitaliers du CHU de Nantes. Avec ces modèles, nous étudions au travers d'une étude comparative l'impact

de différents paramètres sur les performances des modèles de langues pré-entraînés : les stratégies de pré-entraînement de modèles (pré-entraînement de zéro et pré-entraînement poursuivi) et les sources de données (les données biomédicales récoltées sur le web et les données cliniques issues de l'entrepôt de données du CHU de Nantes). Nous introduisons ensuite DrLongformer, un modèle pré-entraîné sur des données biomédicales françaises permettant de traiter des séquences de textes allant jusqu'à 4 096 tokens. Dans ce travail d'adaptation des modèles Longformer, notre objectif est d'évaluer à nouveau les stratégies de pré-entraînement des modèles (pré-entraînement de zéro, pré-entraînement poursuivi et conversion d'un modèle BERT en modèle Longformer) et de vérifier si les résultats observés lors de l'adaptation des modèles BERT sont de nouveau observés avec les modèles Longformer.

Dans le cinquième et dernier chapitre, nous présentons deux cas d'usages des modèles de langue en contexte de recherche clinique. La première section est consacrée au projet GAVROCHE, où nous utilisons les modèles de langue introduits dans cette thèse pour extraire automatiquement des informations cliniques pour phénotyper des patients hospitalisés pour insuffisance cardiaque aiguë. Dans la deuxième section, nous présentons l'application des modèles sur le corpus des déterminants sociaux de santé pour extraire automatiquement ces informations dans les comptes rendus hospitaliers.

Enfin, nous concluons ce manuscrit en revenant sur l'ensemble des contributions de cette thèse et proposons quelques directions de recherche pour poursuivre ces travaux.

Chapitre 1

# Usages du TAL dans les entrepôts de données de santé

Depuis plus de 20 ans, les données relatives à la santé des patients sont systématiquement archivées sous la forme de dossiers médicaux électroniques (ou *Electronic Health Records - EHR* en anglais) dans des bases de données (Meystre et al., 2017; Adler-Milstein et al., 2017). Celles-ci sont créées pour recueillir des données structurées (par exemple, des données biologiques et démographiques) et des données non structurées (par exemple, des comptes rendus textuels sur les hospitalisations ou les consultations). Ces données impliquent de multiples contributeurs : les patients, pour lesquels des données sont collectées lors d'hospitalisations ou de consultations, les soignants qui s'occupent des patients et collectent les données et les établissements de santé qui organisent toute la logistique opérationnelle et financière liée aux soins et à leurs données (Casto, 2013). L'objectif premier de recueil de données est de fournir des soins de qualité aux patients, même si elles peuvent être réemployées dans divers usages secondaires tels que l'optimisation des coûts des soins de santé, la veille sanitaire et la recherche clinique (Meystre et al., 2017). Les données de patients dans la recherche clinique sont limitées en termes de taille d'échantillon, de portée et de suivi longitudinal, notamment dans les essais cliniques ou les registres de maladies. L'usage secondaire des données de santé permet d'augmenter le recrutement des patients dans les essais cliniques (Köpcke and Prokosch, 2014) et d'accéder à une plus grande variété d'informations cliniques pour la recherche (Shah and Khan, 2020; Sarwar et al., 2022). La mise en place d'entrepôts de données (EDS) a permis d'encadrer les usages secondaires de données de santé).

Contrairement aux domaines de la logistique, du marketing et des ventes, le secteur de la santé a été lent à intégrer pleinement les entrepôts de données. Des contraintes de sécurité et de confidentialité liées aux données médicales sont à gérer pour mettre en place un EDS (Hamoud et al., 2018). Selon le pays qui a construit l'EDS, les réglementations relatives aux données médicales peuvent varier et ralentir le processus de

construction (Holmes et al., 2014). Les entrepôts de données font partie du paysage médical depuis des décennies (Gagalova et al., 2020a), en particulier aux États-Unis, où les premiers EDS sont apparus dans les années 90. Dans certains pays, comme la France, les EDS ont été construits plus récemment en raison de contraintes réglementaires. Au niveau institutionnel, l'utilisation des EDS montre que les établissements reconnaissent le potentiel de transformation et la valeur des données générées par leurs activités.

Cette utilisation secondaire des données est facilitée par les avancées technologiques en matière d'IA (Lin et al., 2020b). Parmi les nombreux types de données, les données textuelles renforcent la popularité d'un sous-groupe de méthodes d'IA, le Traitement Automatique des Langues (TAL), qui met en œuvre des algorithmes pouvant fonctionner à une échelle aussi massive que les données textuelles non structurées (Juhn and Liu, 2020). La majorité des informations cliniques sont stockées sous forme de texte non structuré, et le TAL permet d'accéder à ces informations (Névéol et al., 2018; Kim et al., 2019b).

Afin d'étudier l'utilisation du TAL sur les données textuelles des EDS, nous proposons dans cet état de l'art une revue systématique de la littérature. Cette revue a deux objectifs : (1) d'identifier les tâches de TAL qui sont employées pour traiter les documents textuels issus des EDS et (2) identifier pour chaque tâche les méthodes utilisées pour aborder ladite tâche.

Cette revue systématique est organisée en trois étapes :

1. **Méthodologie de la revue systématique.** Dans cette section, nous présentons les critères de sélection des articles, puis les bases de données bibliographiques utilisées ainsi que la stratégie de recherche pour récupérer les articles d'intérêt. Enfin, nous décrivons les informations que nous avons récupérées dans ces articles pour mener cette revue de la littérature.

2. **Résultats de la revue systématique.** Dans cette section, nous décrivons tâche par tâche les résultats.

3. **Discussion des résultats de la revue systématique.** Dans cette section, nous rappelons les résultats principaux et discutons des limitations de cette revue de la littérature.

## 1.1 Méthodologie de la revue systématique

Pour réaliser cette revue systématique de la littérature, nous nous appuyons sur les lignes directrices PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) (Moher et al., 2009).

### 1.1.1 Critères de sélection des articles

Les articles scientifiques retenus pour cette revue ont été inclus selon les critères suivants : (i) articles mentionnant l'utilisation du TAL sur des données provenant d'EDS ; (ii) articles publiés entre 1995 et 2021, et (iii) articles rédigés en anglais. Les critères (ii) et (iii) ont pu être vérifiés automatiquement à l'aide des métadonnées des articles. L'inclusion finale sur le critère (i) a été réalisée manuellement en lisant les titres et les résumés. Lorsque la lecture du titre et du résumé n'était pas suffisant pour statuer de la pertinence de l'article, nous recherchions dans l'article intégral où les mots-clés utilisés dans les requêtes apparaissaient et nous vérifions leur pertinence dans le contexte. Les détails de l'étape de sélection des articles sont décrits dans la figure 1.

### 1.1.2 Bases de données bibliographiques

Trois bases de données bibliographiques ont été requêtées pour récupérer les articles : PubMed, ACL Anthology et Google Scholar. Ces bases de données bibliographiques offrent un panel diversifié d'articles scientifiques. PubMed est spécialisé dans la médecine, la biologie et les sciences de la vie. Son moteur de recherche permet d'élaborer des requêtes combinant les termes *Medical Subject Headings* (MeSH) et le langage naturel. ACL Anthology couvre la littérature publiée dans les conférences relatives à la linguistique computationnelle et le TAL. Google Scholar n'a pas de domaine de spécialité pour les articles qu'il référence et couvre un large éventail de la littérature.

### 1.1.3 Stratégie de recherche

Identifier des articles ayant pour sujet l'application du TAL à des données issues des EDS n'est pas aisé, car cela combine plusieurs désignations. Le terme « entrepôt de données » peut parfois être confondu et retrouvé à travers d'autres termes, tels que « base de données » ou « registre ». En outre, la stratégie d'inclusion des articles repose principalement

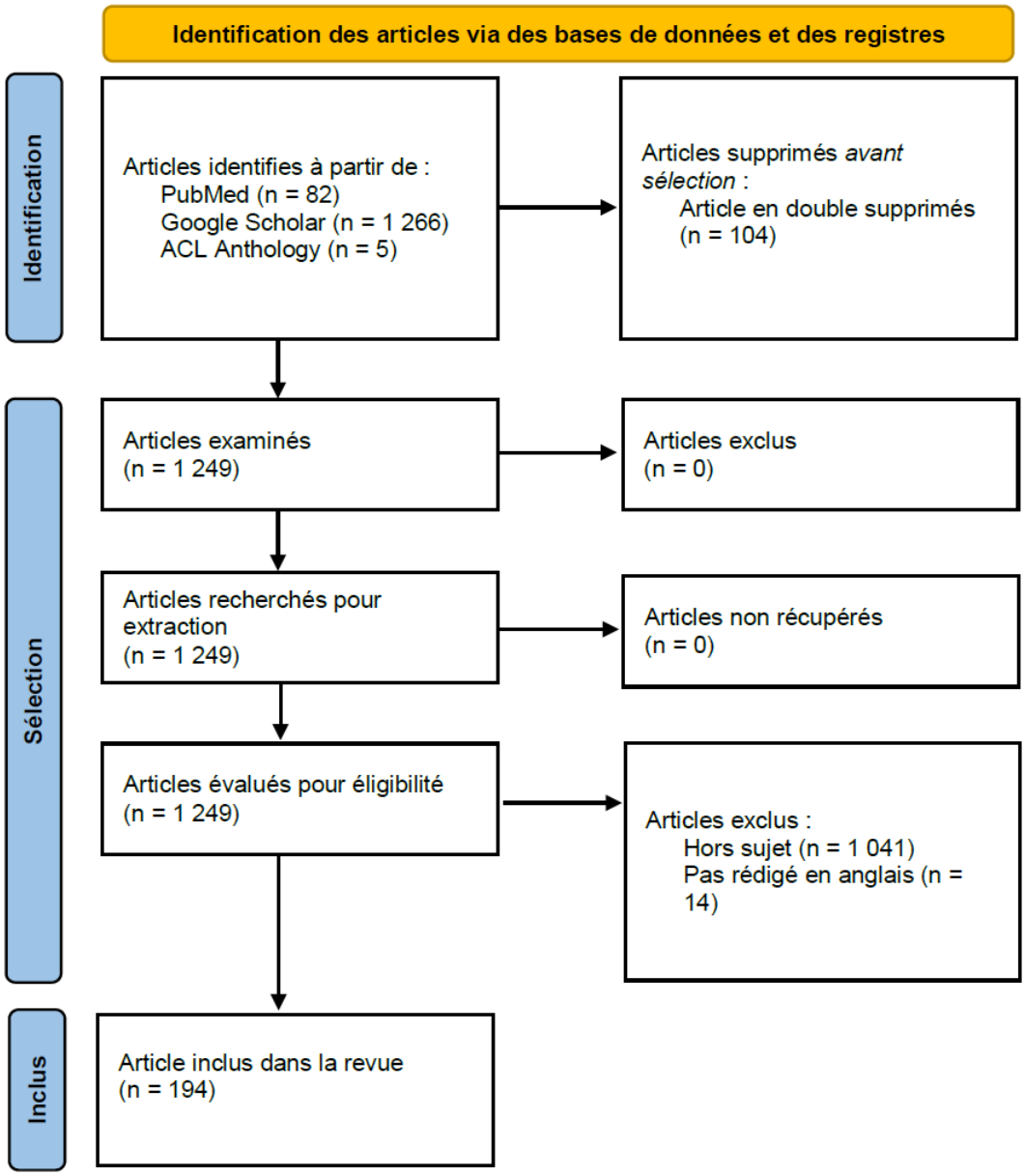


FIGURE 1.1 : Flow-chart de la procédure de sélection des articles pour la revue systématique de la littérature.

sur une lecture du titre et du résumé. Or, la source des données utilisées dans les articles figure souvent uniquement dans l'article intégral et non dans le résumé. La collecte des articles nécessite de multiples requêtes afin d'obtenir une spécificité et une sensibilité élevées.

Afin d'obtenir une sélection représentative d'articles, nous avons utilisé des requêtes

basées sur des mots-clés spécifiques pour chaque sujet d'intérêt, (i) les EDS et (ii) le TAL :

- EDS : *clinical data warehouse*, *biomedical data warehouse*, *health data warehouse*. Les mots-clés choisis pour représenter ce thème correspondent aux noms les plus couramment utilisés pour désigner les EDS.
- TAL : *natural language processing*, *NLP*, *text mining*. Le mot-clé *text mining* est utilisé pour compléter le mot-clé *natural language processing*. Le *text mining* est l'application du TAL la plus utilisée dans le domaine médical. C'est pourquoi le terme *natural language processing* peut parfois être éclipsé par le terme *text mining*.

Plusieurs requêtes ont été construites pour chaque base de données bibliographiques à partir de ces mots clés.

PubMed offre des possibilités de recherche avancées par rapport à ACL Anthology et Google Scholar, facilitant l'intersection entre les deux sujets d'intérêt :

**Requête PubMed**

("data warehousing"[MeSH Terms] OR ("data warehouse")) AND (("clinical") OR ("biomedical") OR ("health"))) AND (("natural language processing") OR ("NLP") OR ("text mining"))

Pour ACL Anthology, nous avons construit trois requêtes. Comme le TAL est le domaine bibliographique couvert par ACL Anthology, les mots-clés liés au TAL ont été exclus des requêtes.

**Requête 1 - ACL Anthology**

"clinical data warehouse"

**Requête 2 - ACL Anthology**

"health data warehouse"

**Requête 3 - ACL Anthology**

"biomedical data warehouse"

Pour Google Scholar, nous avons construit trois requêtes. Pour chacune des requêtes, nous avons lancé une requête similaire supplémentaire en remplaçant *natural language processing* par son acronyme *NLP*. Les résultats de ces requêtes ont été croisés et fusionnés pour être considérés comme une seule requête :

**Requête 1 - Google Scholar**

"clinical data warehouse" "natural language processing"

**Requête 2 - Google Scholar**

"health data warehouse" "natural language processing"

**Requête 3 - Google Scholar**

"biomedical data warehouse" "natural language processing"

Toutes les requêtes ont été réalisées le 23 février 2022. Les articles de PubMed et d'ACL Anthology ont été récupérés après avoir exécuté manuellement les requêtes sur les sites web respectifs de ces bases de données bibliographiques. Les articles de Google Scholar ont été collectés à l'aide d'un logiciel libre[1]. Les résultats des requêtes ont été fusionnés et les doublons ont été supprimés. Les requêtes ne sont pas exhaustives, mais visent plutôt à fournir une sélection représentative d'articles couvrant les sujets d'intérêt. Les synonymes des termes « entrepôts de données » tels que « base de données » ou « registres » n'ont pas été utilisés dans les requêtes afin d'éviter la collecte d'articles non pertinents. De plus, certains articles peuvent également appliquer du TAL à des données provenant d'EDS sans mentionner l'EDS en question et pourraient ne pas être pris en compte dans les requêtes.

## 1.2 Collecte des informations

Les informations suivantes ont été collectées manuellement à partir des articles inclus : (1) les tâches de TAL traitées dans l'article dont la classification est basée sur celle fournie

1. Publish or Perish. `https://harzing.com/resources/publish-or-perish`

par Névéol et al. (2018) ; (2) les méthodes de TAL utilisées pour traiter les tâches ; (3) l'EDS d'où proviennent les données ; (4) la langue des données utilisées dans l'article.

## 1.3 Résultats de la revue systématique

Au total, 1 353 articles (PubMed, n = 82 ; Google Scholar, n = 1 266 ; ACL Anthology, n = 5) ont été identifiés grâce à la stratégie de recherche présentée en section 1.1.3. Après examen du titre et du résumé de chaque article, 1 159 articles ont été exclus pour cause de doublons (n = 104), de problèmes linguistiques (n = 14) ou parce qu'ils n'entraient pas dans le champ d'application de la revue (n = 1 041). Au total, 194 articles répondaient aux critères d'inclusion. Ces 194 articles ont été publiés entre 2002 et 2021, indiquant qu'aucun des articles publiés entre 1995 et 2001 ne répondaient aux critères d'inclusion.

Nous avons identifié dans ces articles les sujets couverts par les recherches publiées sur l'application du TAL aux données des EDS. Nous présentons les résultats des articles examinés par tâche de TAL mentionnée dans les articles. Bien que de nombreux articles traitent de la même tâche de TAL, nous avons décidé de ne pas comparer directement les performances de chaque méthode provenant de différents articles dans cette revue. Les méthodes sont évaluées avec des données différentes, dans des langues différentes, avec des métriques différentes, rendant impossible toute comparaison.

Le tableau 1.1 indique le nombre d'études pour chaque tâche de TAL sur deux périodes temporelles : 2002-2015 et 2016-2021. Le choix de deux périodes est motivé par le changement de paradigme du TAL, qui est passé des méthodes basées sur les connaissances à celles basées sur l'apprentissage automatique, et qui s'accompagne de l'apparition de nouvelles tâches comme la modélisation de la langue.

Les tâches et méthodes de TAL présentées dans le tableau 1.1 ont été appliquées à des données médicales dans différentes langues dont la majorité étaient en anglais (78,86 % des publications ; voir tableau 1.2).

Le tableau 1.3 présente les EDS utilisées dans les articles retenus dans cette revue. En général, les EDS les plus anciens, tels que la base de données cliniques du *Columbia University Medical Center*, la *Mayo Clinic* et le *Partners Healthcare Research Patient Data Repository* sont les EDS qui exploitent le plus les données textuelles et contribuent au développement du TAL.

| **Tâches de TAL (nb. d'articles - %)** | **Méthodes utilisées (nb. d'articles)** | |
|---|---|---|
| | *2002-2015* | *2016-2021* |
| **Extraction d'information** | | |
| Concepts médicaux (37 - 19.1%) | S(14) ML(5) | S(10) ML(11) DL(4) |
| Caractéristiques spécifiques (40 - 20.6%) | S(4) ML(2) | S(22) ML(12) DL(5) |
| Médicaments/effets indésirables (26 - 13.4%) | S(10) ML(3) | S(8) ML(1) DL(5) |
| Observations/Symptômes (8 - 4%) | S(1) ML(1) | S(2) ML(2) DL(4) |
| Relation (1 - 0.5%) | S(1) | - |
| **Classification** | | |
| Phenotypage (38 - 19.4%) | S(7) ML(2) | S(17) ML(12) DL(6) |
| Indexation et codage (7 - 3.6%) | S(3) | S(2) ML(1) DL(1) |
| Topic modeling (3 - 1.5%) | - | S(1) ML(3) |
| Identification des patients (3 - 1.5%) | - | S(1) ML(2) DL(1) |
| **Analyse contextuelle** | | |
| Similarité (6 - 3%) | S(2) | S(1) DL(3) |
| Temporalité (4 - 2%) | S(1) | S(2) |
| Négation (3 - 1.5%) | - | S(2) DL(1) |
| Abbreviation (2 - 1%) | - | S(2) |
| Incertitude (1 - 0.5%) | - | S(1) |
| Expérimentateur (2 - 1%) | - | S(2) |
| **Modélisation du langage (11 - 5.6%)** | - | ML(6) DL(7) |
| **Développement de ressources** | | |
| Corpus et annotation (4 - 2%) | - | ML(1) |
| Lexiques (2 - 1%) | - | S(2) ML(1) |
| **Campagnes d'évaluation (5 - 2.5%)** | S(4) ML(3) | S(1) |
| **Désidentification (2 - 1%)** | S(1) ML(1) | DL(1) |
| **Nettoyage de données (1 - 0.5%)** | - | ML(1) |

Tableau 1.1 : Tâches de TAL renseignées dans les articles sélectionnés. S : méthodes symboliques. ML : apprentissage automatique. DL : apprentissage profond.

| **Langue de données** | **Nombre d'articles** |
|---|---|
| Anglais | 153 |
| Français | 27 |
| Allemand | 9 |
| Coréen | 3 |
| Japonais | 1 |

Tableau 1.2 : Langue des données utilisées dans les articles. Cette information n'a pas été mentionnée dans 2 articles.

| **Entrepôt de données de santé** | **Pays** | **Date** | **# articles** |
|---|---|---|---|
| ACORN data warehouse | USA | 2004 | 1 |
| Albany Medical Center data warehouse | USA | - | 1 |
| Amsterdam University Medical Center | Pays-Bas | - | 1 |
| Antwerp University Hospital (UZA) | Belgique | - | 2 |
| Assistance Publique - Hôpitaux de Paris (AP-HP) | France | 2017 | 6 |
| Boston Children's Hospital data warehouse | USA | - | 2 |
| Carolina Data Warehouse for Health | USA | 2009 | 2 |
| Centre hospitalier universitaire de Sherbrooke | Canada | - | 1 |
| Columbia University Medical Center clinical data warehouse | USA | 1994 | 30 |
| Enterprise Clinical Research Data Warehouse Hannover Medical School | Allemagne | 2011 | 1 |
| Entrepôt de données du CHU de Tours | France | 2019 | 1 |
| Georges Pompidou European Hospital | France | 2008 | 4 |
| Houston METEOR | USA | 2012 | 5 |
| Indiana Network for Patient Care | USA | 1998 | 1 |
| Intermountain Healthcare enterprise-wide data warehouse | USA | 1998 | 3 |
| Kaiser Permanente Southern California | USA | - | 1 |
| Korian (groupe privé) | France | 2010 | 1 |
| Loyola University Medical Center | USA | 2003 | 4 |
| Mayo Clinic | USA | 2005 | 24 |
| McGill University Health Centre | Canada | 2019 | 2 |
| Medical University of South Carolina Research Data Warehouse | USA | 2013 | 3 |
| Mount Sinai Hospital Data Warehouse | USA | 2011 | 4 |
| Northwestern Enterprise Data Warehouse | USA | 2007 | 2 |
| Osaka University Medical Hospital | Japon | - | 1 |
| Paris Necker Children's Hospital | France | 2017 | 7 |
| Partners Healthcare Research Patient Data Repository | USA | 2002 | 13 |
| Rennes University Hospital | France | 2018 | 2 |
| Rouen University Hospital | France | 2019 | 3 |
| Samsung Medical Center | Corée du Sud | - | 3 |
| Seoul National University Hospital (SUPREME) | Corée du Sud | - | 2 |
| Stanford Medicine Research Data Repository (STARR) | USA | 2008 | 8 |
| STRIDE Clinical Data Warehouse | USA | 2008 | 10 |
| University Hospital of Erlangen | Allemagne | 2009 | 1 |
| University Hospital of Bordeaux | France | 2019 | 2 |
| University Hospital of Würzburg | Allemagne | - | 2 |
| University of Arizona, Banner University Medical Center | USA | - | 1 |
| University of Arkansas for Medical Sciences Epic MER System | USA | 2019 | 1 |
| University of California, Irvine medical center clinical data warehouse | USA | 2011 | 1 |
| University of Florida Health Integrated Data Repository | USA | - | 3 |
| UT-Physicians | USA | - | 7 |
| UW Health, School of Medicine and Public Health of the University of Wisconsin-Madison | USA | - | 1 |
| VA Corporate Data Warehouse | USA | 2006 | 9 |
| Vanderbilt University Medical Center Synthetic | USA | 2014 | 3 |

TABLEAU 1.3 : Entrepôts de données de santé exploités dans les articles.

### 1.3.1 Extraction d'information

Le tableau 1.1 montre que l'extraction d'informations est l'une des tâches de TAL les plus étudiées dans le domaine clinique. Une tâche type de l'extraction d'informations est la reconnaissance d'entités nommées (REN) qui se concentre dans le domaine clinique sur des entités telles que les informations de santé identifiantes pour désidentifier ou pseudonymiser les documents cliniques (Ferrández et al., 2013; Yang et al., 2019) et les concepts cliniques, dont les maladies (Ashish et al., 2014; Osborne et al., 2016; Walsh et al., 2017; Hong et al., 2017; Afzal et al., 2017; Hunter-Zinck et al., 2019; Lerner et al., 2020b), les observations et les symptômes (Pham et al., 2014; Chokshi et al., 2017; Patel et al., 2017; Fiebeck et al., 2018; Lerner et al., 2020b; Neuraz et al., 2020; Min et al., 2022), les médicaments (Gold et al., 2008; LePendu et al., 2012b,a; Jung et al., 2013; Weeks et al., 2020; Geva et al., 2020; Lerner et al., 2020b; Neuraz et al., 2020; Lerner et al., 2020a; Hoertel et al., 2021; Caliskan et al., 2021; Chouchana et al., 2021; Jouffroy et al., 2021) et leurs informations associées : dosage, fréquence et durée de traitement (Neuraz et al., 2020; Lerner et al., 2020a; Weeks et al., 2020; Hoertel et al., 2021; Caliskan et al., 2021; Chouchana et al., 2021; Jouffroy et al., 2021) ou effets indésirables (LePendu et al., 2012b,a; Leeper et al., 2013; LePendu et al., 2013; Wright et al., 2013; Jung et al., 2015; Wang et al., 2015a; Rochefort et al., 2015, 2017; Shimai et al., 2018; Geva et al., 2020). Les concepts cliniques peuvent être reliés à des ressources terminologiques ou des ontologies comme l'UMLS (*Unified Medical Language System*) (Lowe et al., 2009; Jonnalagadda et al., 2012; Campillo-Gimenez et al., 2012; Jonnalagadda et al., 2013; Wang et al., 2015a; Singh et al., 2016; Chase et al., 2017; Zhou et al., 2019; Hunter-Zinck et al., 2019), la SNOMED-CT (*Systematized Nomenclature Of MEDicine - Clinical Terms*) (Melton et al., 2006; Wang et al., 2008; Lowe et al., 2009) ou la CIM-9 (Classification internationale des maladies) (Perotte et al., 2011).

Certains systèmes de TAL sont largement utilisés pour extraire, structurer et coder les informations cliniques contenues dans les comptes rendus cliniques en anglais. Des études décrivent comment utiliser **MedLEE** (*Medical Language Extraction and Encoding System*) pour l'extraction de concepts cliniques (Chuang et al., 2002; Cao et al., 2005; Melton et al., 2006; Wang et al., 2008; Van Vleck et al., 2008; Li et al., 2008; Chen et al., 2008; Wang et al., 2009; Carlo et al., 2010; Wang et al., 2010; Overby et al., 2013; Chase et al., 2017) ou de médicaments (Li et al., 2011; Harpaz et al., 2013; Malec et al., 2021). L'extraction et la liaison des informations cliniques avec l'*UMLS* ont également été réalisées avec des outils comme **cTAKES** (*clinical Text Analysis and Knowledge Ex-*

*traction System*) (Walsh et al., 2017; Zhong et al., 2018; Afshar et al., 2019a; Raja et al., 2019; Afshar et al., 2019b; To et al., 2020; Sharma et al., 2020; Geva et al., 2020; Zong et al., 2021), **MetaMap** (Singh et al., 2016; Osborne et al., 2016; Zhou et al., 2019; Harris et al., 2020), **MedTagger** (Liu et al., 2012a; Afzal et al., 2017; Moon et al., 2018; Wang et al., 2019c; Zhao et al., 2021, 2022) et **NCBO** (*National Center for Biomedical Ontology*) **Annotator** (LePendu et al., 2012b,a; Leeper et al., 2013; Jung et al., 2013, 2015; Wang et al., 2015a). Les concepts extraits peuvent également être reliés à d'autres ontologies standardisées comme la SNOMED-CT (Melton et al., 2006). Pour les langues autres que l'anglais, peu d'outils similaires ont été proposés. Par exemple, l'outil **AHD NLP** (Caliskan et al., 2021) a été évalué sur une tâche d'extraction de médicaments sur des documents cliniques en allemand.

Les informations cliniques recherchées dans les documents ne sont pas toujours reliées à des nomenclatures. Dans certains cas d'usages spécifiques, il est nécessaire de développer des systèmes adaptés aux informations recherchées et plusieurs approches sont possibles. Les méthodes basées sur des règles encodent des dictionnaires et des terminologies pour faire correspondre des termes et des concepts dans les textes cliniques (Gold et al., 2008; LePendu et al., 2013; Hong et al., 2017; Hunter-Zinck et al., 2019; Kshatriya et al., 2020; Lerner et al., 2020b; Weeks et al., 2020; Chouchana et al., 2021). Les méthodes d'apprentissage automatique tirent parti des connaissances cliniques présentes dans la grande quantité de données des EDS. Selon la période temporelle, les méthodes employées reflètent la tendance des méthodes état de l'art. Les champs aléatoires conditionnels (*Conditional Random Fields - CRF*) ont été utilisés pour extraire des concepts cliniques (Jonnalagadda et al., 2012, 2013) ou désidentifier des documents cliniques (Ferrández et al., 2013). L'allocation de Dirichlet latent supervisée hiérarchiquement (*Hierarchically Supervised Latent Dirichlet Allocation - HSLDA*) a été appliquée à des résumés de sortie d'hôpital afin de prédire les codes CIM-9 (Perotte et al., 2011). Les approches d'apprentissage profond telles que les *bidirectional long short-term memory* (BiLSTM-CRF) (Lerner et al., 2020a; Chouchana et al., 2021; Jouffroy et al., 2021), les *recurrent neural network grammar* (RNNG) (Lerner et al., 2020a) ont été employés pour de l'extraction d'entités médicales dans des textes cliniques en français. Chokshi et al. (2017) ont comparé les sacs de mots (*bag-of-words*) combiné à une machine à vecteurs de support (*Support Vector Machine - SVM*) à deux modèles de réseaux neuronaux : les réseaux neuronaux convolutifs (*Convolutional Neural Network - CNN*) et le modèle d'attention neuronal (*Neural Attention Model - NAM*), tous deux avec des plongements Word2Vec comme représentations des mots en

entrée. Les précisions des modèles CNN et NAM étaient relativement égales, mais plus élevées que celles du modèle SVM. Lerner et al. (2020b) ont comparé trois systèmes de reconnaissance d'entités nommées cliniques : (i) un système basé sur les terminologies UMLS et SNOMED, (ii) un système biGRU-CRF et (iii) un système hybride utilisant la prédiction du premier système comme caractéristique pour le deuxième système biGRU-CRF. Yang et al. (2019) ont désidentifié des informations de santé identifiantes et sensibles à l'aide d'un modèle LSTM-CRF.

Des modèles plus récents, basés sur l'architecture *Transformer* (Vaswani et al., 2017), sont également employées pour extraire des concepts médicaux dans des textes cliniques. Neuraz et al. (2020) ont utilisé une couche BiLSTM-CRF au-dessus d'une représentation vectorielle calculée à partir d'un modèle BERT multilingue et ont appliqué ce modèle à des données en français. BERT et RoBERTa ont été entraînés pour extraire des déterminants sociaux de santé (DSS) à partir de comptes rendus cliniques (Yu et al., 2021b). Certains travaux associent un modèle de langage à des techniques simples de *pattern-matching*. Par exemple, Jouffroy et al. (2021) ont proposé une approche hybride pour l'extraction de médicaments à partir d'un texte clinique français en combinant des expressions régulières pour pré-annoter le texte avec des plongements de mots contextualisés (ELMo) qui sont introduits dans un réseau neuronal récurrent profond (BiLSTM-CRF).

Certaines études portent sur l'extraction d'informations cliniques spécifiques, telles que la densité osseuse (Wang et al., 2019b), les mentions BRCA1/2 (Zhao et al., 2021), les prédicteurs et le moment de changement de mode de vie pour les patients souffrant d'hypertension (Shoenbill et al., 2020), la caractérisation d'une imagerie positive dans les rapports de radiologie (Sharperson et al., 2021), la classification de Banff (Zubke et al., 2020), l'infection du site chirurgical (Ciofi Degli Atti et al., 2020), le niveau BI-RADS 3 (Cochon et al., 2020; Lacson et al., 2020), la toxicité de la chimiothérapie (Rogier et al., 2021), les signes vitaux (Genes et al., 2013), la résection transurétrale d'une tumeur de la vessie (Glaser et al., 2018), l'utilisation de statines (Riestenberg et al., 2019), les génotypes HLA (Lee et al., 2020b), l'épisode de soins non planifié (Tamang et al., 2015), le tabagisme (Bae et al., 2021; Yang et al., 2020a), la gammapathie monoclonale (Ryu and Zimolzak, 2020), les fractures spécifiques au site squelettique (Wang et al., 2019d) ou encore les déterminants sociaux de santé (Stemerman et al., 2021). Les méthodes utilisées pour extraire ces informations sont fondées sur des règles (Genes et al., 2013; Tamang et al., 2015; Glaser et al., 2018; Wang et al., 2019d; Lee et al., 2020b; Yang et al., 2020a; Zubke et al., 2020; Ciofi Degli Atti et al., 2020; Cochon et al., 2020; Lacson et al., 2020;

Rogier et al., 2021; Sharperson et al., 2021), des approches statistiques (Wang et al., 2019b; Shoenbill et al., 2020) ou une combinaison des deux (Ryu and Zimolzak, 2020; Zhao et al., 2021). De multiples informations sur les patients peuvent être extraites de textes cliniques pour être ensuite exploitées dans des études rétrospectives (Sehdev et al., 2018). Ansoborlo et al. (2021) ont extrait 52 informations biocliniques de comptes rendus de réunion d'équipe pluridisciplinaire française en appliquant des expressions régulières ou une méthode de classification bayésienne.

L'extraction d'informations à partir d'un texte clinique peut également se faire sous la forme d'une tâche de prédiction, comme des indicateurs de réadmission, de durée de soin. Parmi les indicateurs prédits, on retrouve par exemple la durée du séjour à l'hôpital (Chrusciel et al., 2021), la probabilité d'admission à l'unité de soins intensifs en neurosciences (Klang et al., 2021), le risque de réadmission à 30 jours chez les patients souffrant d'insuffisance cardiaque (Golas et al., 2018), ou la mesure de la qualité de la documentation sur le toucher rectal (Bozkurt et al., 2019). Pour les indicateurs de risque de maladies ou de pathologies, on retrouve le VIH (Feller et al., 2018b; Duthe et al., 2021), le cancer du pancréas (Chen et al., 2020c), les escarres (Luther et al., 2017), l'insuffisance rénale chronique (Perotte et al., 2015) et le cancer du sein (He et al., 2019). La prédiction de ces informations cliniques peut être réalisée à l'aide de règles (Chrusciel et al., 2021; Duthe et al., 2021), de méthodes d'apprentissage automatique telles que l'allocation de Dirichlet latente (LDA) (Perotte et al., 2015; Chrusciel et al., 2021), ou d'une combinaison des deux (Luther et al., 2017; Chen et al., 2020c).

### 1.3.2 Analyse contextuelle

Les informations extraites des comptes rendus cliniques nécessitent parfois l'analyse de leur contexte d'occurrence pour juger de leur pertinence. Parmi ces informations contextuelles, nous retrouvons la négation, la temporalité, l'incertitude ou l'expérimentateur (c'est-à-dire déterminer si l'information identifiée est liée au patient ou à un tiers, tel qu'un membre de la famille). Des règles sous forme de patrons permettent de détecter les informations contextuelles dans les textes cliniques (Garcelon et al., 2017a; Mirzapour et al., 2021; Klappe et al., 2021). Bien que ces méthodes offrent de bons résultats (0,90 de F-mesure en moyenne), elles nécessitent des lexiques souvent constitués par des experts. Une adaptation est souvent nécessaire pour des cas d'usages spécifiques. La temporalité ont été étudiée par Liu et al. (2012b) pour distinguer les effets indésirables des médicaments des indications dans les textes cliniques. Zhou et al. (2006) décrivent une structure

de contraintes temporelles construite à partir d'expressions temporelles dans les résumés de sortie d'hôpital pour modéliser ces expressions. Dans le domaine clinique, de nombreuses expressions temporelles présentent des caractéristiques uniques et cette structure offre une couverture complète pour l'encodage de ces expressions. Les abréviations médicales les plus courantes ont été étudiées dans les textes cliniques français par Cossin et al. (2021) et anglais par Moon et al. (2017). Les méthodes récentes de plongements contextualisées, tels que BERT, ont facilité l'étude de la détection de la négation (Lin et al., 2020a) et de la similarité des textes (Mahajan et al., 2020; Li et al., 2021a). La similarité des textes est également étudiée pour identifier des concepts sémantiquement similaires (Pivovarov and Elhadad, 2012), des patients similaires (Garcelon et al., 2017b) ou pour détecter la redondance dans les textes cliniques (Cohen et al., 2013; Mutinda et al., 2020).

### 1.3.3 Classification

L'identification des patients est un élément-clé de la recherche clinique pour construire les populations des études. Le TAL peut améliorer l'interrogation et l'indexation des patients et de leurs données dans les EDS. Zhu et al. (2014) ont abordé l'expansion des requêtes sur la base d'un vaste corpus clinique pour résoudre les problèmes de polysémie, de synonymie et d'hyponymie dans les textes cliniques. L'expansion des requêtes se décline en trois méthodes principales : à l'aide de synonymes, de classes sémantiques construits à partir des textes ou de dérivés morphologiques (Zeng et al., 2012). Un algorithme de recherche automatisé pour identifier les complications postopératoires a été évalué par (Tien et al., 2015). Un EDS sémantique a été conçu pour aider les professionnels de santé à présélectionner les patients éligibles aux essais cliniques (Lelong et al., 2019; Pressat-Laffouilhère et al., 2022). Certaines études s'appuient sur Dr Warehouse, un entrepôt de données biomédicales, développé à l'hôpital Necker-Enfants malades, basé sur les comptes rendus cliniques. Cet entrepôt de données est utilisé pour explorer, à l'aide de la fréquence et du TF-IDF, l'association entre les phénotypes cliniques et les maladies rares telles que la variante KCNA2 dans les syndromes neurodéveloppementaux (Hully et al., 2021), le syndrome de Dravet (Lo Barco et al., 2021), la ciliopathie (Chen et al., 2019c) et d'autres maladies rares (Garcelon et al., 2018).

En aval de l'interrogation des EDS, le TAL peut être appliqué aux documents cliniques pour identifier les patients ou les documents d'intérêt lorsque les méthodes de classification offertes par les outils intégrés dans les EDS ne sont pas suffisamment précis.

L'identification des patients peut se faire à l'aide de différentes méthodes telles que des règles, en utilisant des termes liés aux critères d'inclusions (Stephen et al., 2003; Haerian et al., 2012; Nigwekar et al., 2014; Ahmed et al., 2014; Yahi and Tatonetti, 2015; Evans et al., 2016; Upadhyaya et al., 2017; Redman et al., 2017; Krebs et al., 2018; Afzal et al., 2018; Zhu et al., 2019; Bastarache et al., 2019; Hoffman et al., 2018; Meystre et al., 2019; Wu et al., 2022), des méthodes d'apprentissage automatique (Carter et al., 2015; Shao et al., 2019; Alba et al., 2021) ou une combinaison des deux (Chase et al., 2010; Kim et al., 2019a; Ling et al., 2019; Bozkurt et al., 2020; Ferté et al., 2021). Une combinaison de données structurées et non structurées en allemand a été utilisée par Scheurwegs et al. (2016) pour attribuer des codes cliniques aux séjours des patients. Chen et al. (2017) et Li et al. (2021b) ont appliqué un modèle LDA à des documents cliniques pour la modélisation des sujets (*topic modeling*). Agarwal et al. (2016) ont détaillé un modèle de régression logistique des phénotypes appris sur des données étiquetées bruitées.

### 1.3.4 Modélisation du langage

Les méthodes récentes de plongements de mots tirent parti de la grande quantité de données stockées dans les EDS pour apprendre des représentations sémantiques à partir des textes cliniques. Ces méthodes permettent, par exemple, de mesurer la similarité des termes dans l'espace de plongement (Magnani et al., 2021; De Freitas et al., 2021). Parmi ces méthodes, les modèles *Transformers*, tels que BERT (Devlin et al., 2019), peuvent être affinés (*fine-tuned*) sur plusieurs tâches, notamment la classification de texte pour faire correspondre le titre du document à la terminologie LOINC (Zuo et al., 2020) et l'étiquetage de séquences pour détecter et estimer l'emplacement des anomalies dans les TEP scans (Eyuboglu et al., 2021). De même, la classification des codes CIM, obtenu à partir de méthodes de vectorisation, permet de structurer les textes cliniques (Zhan et al., 2021; Lee et al., 2023).

Certaines études ont évalué l'efficacité des modèles de plongements de mots sur plusieurs tâches. Word2Vec a été appliqué dans de nombreuses études à des fins différentes, notamment pour évaluer l'utilisation de la scintigraphie osseuse chez les patients atteints de cancer de la prostate à l'aide d'un CNN (Coquet et al., 2019), pour dépister et diagnostiquer le cancer du sein à l'aide d'une architecture d'apprentissage profond (He et al., 2017), pour extraire des caractéristiques et prédire le risque de transplantation hépatocellulaire chez les patients atteints de cancer du foie à l'aide d'un réseau CapsNet (He et al., 2021b) ou pour apprendre avec un CNN le statut d'éligibilité des patients pour

participer à des études de cohorte (Chen et al., 2019a). Lee et al. (2020a) ont proposé une plateforme d'apprentissage de représentation de graphes unifié basé sur les réseaux convolutifs de graphes (*Graph Convolutional Networks - GCN*) et les LSTM pour construire une représentation sous forme de graphe des entités médicales contenues dans les documents cliniques. Dligach et al. (2019) ont développé un encodeur de texte clinique pour le phénotypage. Des expériences ont été menées avec des DAN et des CNN pour construire cet encodeur de texte.

### 1.3.5 Développement de ressources et campagnes d'évaluation

De nombreuses méthodes de TAL nécessitent des ressources spécifiques au domaine clinique pour être déployées. Les données issues des EDS combinées à l'expertise clinique permettent de développer des ressources telles que des schémas d'annotation (Van Vleck et al., 2007; Feller et al., 2018a; Roberts et al., 2018), des lexiques (Song et al., 2021), des ontologies (Loda et al., 2019) ou des plateformes pour valider les résultats des systèmes de TAL (Escudié et al., 2015).

Les efforts récents de la communauté internationale ont conduit à l'introduction de tâches partagées (*shared tasks*) à partir de documents cliniques provenant des EDS. Le défi de l'obésité i2b2 s'est concentré sur l'obésité et ses 15 comorbidités les plus courantes grâce à une tâche de classification à classes multiples et à étiquettes multiples (Uzuner, 2009; Solt et al., 2009). Un autre défi i2b2 organisé en 2009 portait sur l'extraction d'informations de médicaments à partir de textes cliniques (Uzuner et al., 2010; Patrick et al., 2011). Trois tâches ont été proposées lors du quatrième défi *i2b2/VA shared-task and workshop challenge* : (i) l'extraction de problèmes médicaux, d'examens et de traitements ; (ii) la classification des affirmations concernant des problèmes médicaux ; et (iii) la classification de relations entre des paires de concepts apparaissant dans une même phrase, où au moins un concept d'une paire est un problème médical (Patrick et al., 2011). Ces tâches partagées i2b2 exploitent des comptes rendus de sortie désidentifiées provenant du *Partners Healthcare Research Patient Data Repository*. L'édition 2018 de l'atelier *National NLP Clinical Challenges (n2c2)* a présenté une tâche de sélection de cohorte pour les essais cliniques (Chen et al., 2019b).

## 1.4 Discussion des résultats

Les EDS sont de plus en plus répandus et adoptés dans de nombreux pays, ce qui permet au TAL clinique de se développer. Cette revue systématique de la littérature montre que l'utilisation du TAL sur les données des EDS est principalement consacrée à l'extraction d'informations à partir des textes cliniques et à l'identification de populations de patients. En fonction de la tâche cible, différentes méthodes peuvent être utilisées, allant des méthodes symboliques aux méthodes d'apprentissage automatique et d'apprentissage profond pour les plus récentes. Les méthodes symboliques et linguistiques sont encore largement utilisées ces dernières années, malgré la prépondérance des approches d'apprentissage profond qui donnent de très bons résultats pour une majorité de tâches. Cela indique que certaines tâches peuvent être partiellement réalisées avec des techniques classiques de TAL, telles que les expressions régulières ou la recherche de motifs en exploitant des lexiques spécialisés, comme des listes de médicaments ou des terminologies. Les outils d'extraction d'informations existants, tels que cTAKES, MedLee ou MetaMap, offrent une prise en main facile et des résultats satisfaisants et sont donc souvent utilisés pour traiter les textes cliniques en anglais.

Il est intéressant de noter que le nombre de langues présentées dans cette revue systématique concerne seulement cinq langues : anglais, français, allemand, coréen, japonais. Une possible explication serait : (1) les EDS ne sont pas cités comme source de données dans les articles, ce qui constitue un biais lié aux requêtes ; (2) les EDS sont opérationnels dans d'autres pays, mais le TAL n'est pas encore utilisé sur ces données ; (3) les EDS ne sont pas encore adoptés dans tous les pays.

Les EDS les plus anciens sont ceux avec le plus de publications. L'utilisation du TAL n'est donc pas un événement ponctuel, mais a plutôt vocation à s'installer sur le long terme et à contribuer à une amélioration continue de la qualité des données mises à disposition dans les EDS.

## 1.5 Limitations

Les tâches de TAL identifiées dans cette revue ne couvrent qu'une partie que celles existantes dans le domaine général. Ces tâches reflètent globalement les principaux besoins de la recherche clinique, tels que l'identification de la population pour une étude et l'extraction d'informations cliniques pour une population définie. Les autres tâches, telles

que l'analyse contextuelle et la modélisation du langage, sont largement étudiées dans le domaine général du TAL mais sont moins populaires dans le domaine clinique. Ces dernières années, les approches fondées sur l'architecture *Transformers* constituent les méthodes état de l'art dans la plupart des tâches de TAL. Notre revue systématique indique que ces méthodes n'ont pas encore pleinement imprégné le domaine clinique. Il existe donc un écart entre une méthode répandue dans le domaine général du TAL et son appropriation dans un domaine spécialisé, tel que le domaine clinique.

Notre revue systématique se concentre sur deux sujets très spécifiques issus de domaines émergents, le TAL clinique et les EDS. Ce sujet double implique la consultation de plusieurs bases de données bibliographiques et l'agrégation de plusieurs requêtes afin d'obtenir une bonne couverture de la littérature. Certaines bases de données bibliographiques couvrent un plus large éventail d'articles et incluent des articles déjà présents dans d'autres bases plus spécialisées. Pour éviter les doublons, nous avons utilisé en priorité les bases de données bibliographiques les plus fournies et les plus complètes, Google Scholar et PubMed. Cela introduit un biais d'exhaustivité car des articles pertinents pourraient manquer dans les bases de données bibliographiques sélectionnées et être présents dans d'autres bases de données que nous n'avons pas utilisées dans cette revue, telles que Scopus, Web of Science ou Embase. Un autre biais d'exhaustivité découle de la recherche par mots-clés dans les bases de données bibliographiques est présent. En effet, un concept donné peut être exprimé de plusieurs manières en langage naturel, au moyen de différents mots-clés. Le choix des mots-clés est crucial pour obtenir à la fois une spécificité et une sensibilité élevées, même si les mots-clés sélectionnés sont recherchés dans l'ensemble de l'article. Dans cette revue, nous avons utilisé des mots-clés très larges pour obtenir la plus grande sensibilité, mais au détriment de la spécificité (194 articles pertinents parmi les 1,353 articles résultant des requêtes).

## 1.6 Conclusion

Bien que le TAL soit de plus en plus populaire, il reste encore peu développé dans les domaines de spécialité comme le domaine clinique. Les EDS sont encore sous-exploités pour les études médicales spécifiques et plus globalement pour la recherche en TAL. Les EDS sont rarement en accès libre pour la recherche en TAL en raison de la confidentialité des données patients. Une première étape pour surmonter partiellement les contraintes liées à la protection de la vie privée pourrait consister à travailler sur des données désiden-

tifiées, pseudonymisées ou anonymisées provenant des EDS. Cette approche a été utilisée dans le cadre de certaines tâches (*shared tasks*) récentes (Uzuner, 2009; Solt et al., 2009; Uzuner et al., 2010; Patrick et al., 2011). Les *shared tasks* sont cruciales pour faire progresser la recherche sur les domaines spécialisés. Cependant, elles sont peu nombreuses, en particulier pour les langues autres que l'anglais (Névéol et al., 2018).

La tendance globale est à la structuration et à l'interopérabilité des données cliniques. Cependant, la finesse du raisonnement médical reste exprimée dans des comptes rendus textuels et ne font pas l'objet de structuration. Les données structurées ou semi-structurées stockées dans le EDS fournissent des informations sur le suivi des patients et peuvent constituer des ressources utiles pour développer ou améliorer les systèmes de TAL.

L'application de méthodes de TAL sur des documents cliniques provenant d'EDS français représentent à peine 14 % des articles rapportés dans cette revue de la littérature. Néanmoins, ces usages sont de plus en plus présents ces dernières années, notamment grâce aux nombreux EDS s'installant dans le paysage hospitalier français. Dans ce manuscrit, nous présentons deux nouvelles applications du TAL sur des données cliniques issues de l'EDS du CHU de Nantes. Le premier cas d'usage est l'extraction d'informations pour phénotyper des patients hospitalisés pour insuffisance cardiaque aiguë. Le deuxième cas d'usage porte sur l'extraction de déterminants sociaux de santé dans les documents cliniques. Ces deux premiers cas d'usage sont présentés en deux chapitres. L'annotation des corpus est présentée dans le chapitre 3 tandis que le développement des systèmes d'extraction d'informations sur chacun des corpus est présenté dans le chapitre 5.

Chapitre 2

# Représentations sémantiques des mots et évaluation extrinsèque

Le TAL que ce soit dans le domaine général ou dans le domaine médical, élabore et exploite des méthodes statistiques et en particulier les méthodes d'apprentissage automatique ou d'apprentissage profond. L'utilisation de telles méthodes nécessite une représentation du texte adaptée à ces outils, généralement sous forme vectorielle. Diverses formes de représentations vectorielles du texte ont été proposées, allant des représentations discrètes aux représentations numériques.

L'évolution de ces méthodes statistiques s'est accompagnée conjointement d'une évolution des représentations vectorielles des mots avec des représentations de plus en plus performantes pour incorporer la sémantique des mots.

Dans ce chapitre, nous décrivons de manière chronologique les méthodes permettant de représenter les mots, allant des représentations discrètes (les sacs de mots), aux représentations continues et contextualisées permettant de représenter les mots en fonction de leur contexte, notamment à travers les modèles de langues basés sur l'architecture Transformer. Nous présentons également les méthodes permettant d'adapter les modèles de langue à des domaines spécialisés.

## 2.1 Représentations discrètes : sac de mots

Les sacs de mots s'inspirent de la sémantique distributionnelle telle qu'introduit par Zellig S. Harris (Harris, 1954). L'idée sous-jacente des sacs de mots repose sur la notion que l'information essentielle dans un texte peut être capturée en considérant simplement les occurrences des mots qui le composent, tout en ignorant leur ordre et leur structure grammaticale. Le texte est donc représenté par un vecteur de la même taille que le vocabulaire. La figure 2.1 propose un exemple de sac de mots pour une phrase simple.

Les sacs de mots avec les occurrences des mots, comme présenté ci-dessus, permettent

Texte $T_1$ = « Une patiente présente une fièvre élevée »

Vocabulaire = {une, patiente, présente, fièvre, élevée}

| | une | patiente | présente | fièvre | élevée |
|---|---|---|---|---|---|
| $T_1 =$ | | | | | |
| $S(T_1) =$ | 2 | 1 | 1 | 1 | 1 |

FIGURE 2.1 : Exemple de sac de mots pour une phrase simple où $T_1$ est un texte et $S(T_1)$ est un vecteur .

d'avoir une représentation vectorielle d'un texte ou d'un document, mais ne permettent pas de fournir une représentation propre à chaque mot. En revanche, des représentations vectorielles pour les mots peuvent être obtenues en se basant sur les cooccurrences des mots plutôt que les occurrences. La représentation vectorielle d'un mot $w$ est définie en fonction des mots $c$ qui l'entourent dans une fenêtre contextuelle $f$. De façon plus formelle, un mot $c$ cooccurre avec un mot $w$ si $c \in [w_{-f}, -1] \cup [1, w_f]$, où $f$ correspond à la taille de la fenêtre de contexte et $[w_{-f}, -1] \cup [1, w_f]$ aux mots présents dans la fenêtre contextuelle de $w$.

Ce calcul est effectué pour toutes les occurrences d'un mot dans le corpus et permet d'obtenir un vecteur de contexte pour chaque mot du vocabulaire $\mathcal{V}$. Chaque dimension du vecteur correspond donc à un mot du vocabulaire et au nombre de cooccurrences avec le mot initial. L'ensemble des vecteurs des mots du vocabulaire peut ensuite être concaténé pour former une matrice de cooccurrences de taille $\mathcal{V} \times \mathcal{V}$, comme illustré dans la figure 2.2.

Les sacs de mots sont simples à mettre en œuvre et permettent de représenter les mots efficacement, en particulier lorsqu'on travaille avec de grands corpus de textes. Un autre avantage est l'interprétabilité des sacs de mots. En effet, les dimensions qui composent les vecteurs correspondent aux mots du vocabulaire et permettent de donner une représentation précise et directement compréhensible par l'humain. Cependant, les sacs de mots présentent plusieurs limites. La première est la faible qualité des représentations des mots peu fréquents. Plus un mot apparaît dans le corpus, plus ce mot aura des contextes variés dans lesquels il apparaît, ce qui favorisera une représentation vectorielle précise et dense. À l'inverse, la représentation d'un mot peu fréquent sera moins fiable, car moins de cooccurrences, donnant ainsi un vecteur creux avec beaucoup de dimensions à 0.

Une première intuition pour améliorer la qualité des représentations vectorielles serait

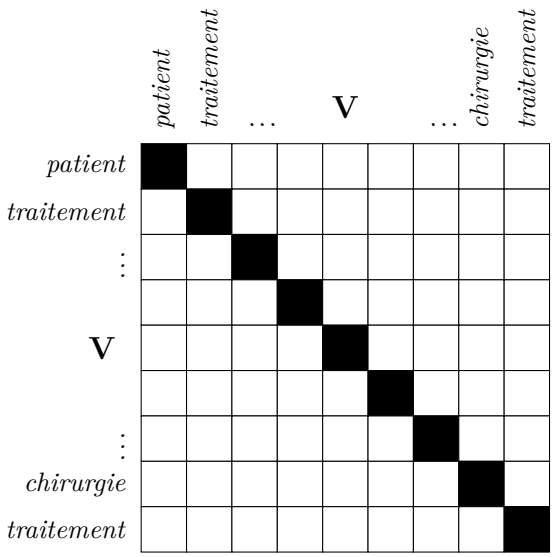


FIGURE 2.2 : Représentation d'une matrice de cooccurrences pour un corpus de taille de vocabulaire $V$.

d'augmenter la taille du corpus. Pourtant, augmenter la taille du corpus fait également augmenter la taille du vocabulaire et donc le nombre de dimensions des vecteurs, ce qui peut poser problème. En effet, la taille du vocabulaire est aussi un des inconvénients des sacs de mots. Sur des corpus de plusieurs millions de mots, le vocabulaire peut facilement atteindre une taille de plusieurs centaines de milliers de mots uniques. Cela a pour résultat des vecteurs creux car la majorité des mots apparaissent peu et ne cooccurrent pas avec d'autres.

L'association entre la grande dimensionnalité des vecteurs et leur caractère épars rend difficile l'application des méthodes d'apprentissage profond qui sont optimisés pour des vecteurs denses et de faible dimension. Pour remédier à cette difficulté, des travaux se sont intéressés à représenter les mots dans des espaces vectorielles de dimension fixe : les représentations continues, que nous présentons dans la section suivante.

## 2.2 Représentations continues et statiques

Pour pallier les problèmes de matrices creuses et de hautes dimensions, de nouvelles des méthodes dites de plongements de mots (ou *embeddings* en anglais) ont été introduites. Celles-ci sont fondées sur des représentations vectorielles continues de mots où chaque mot est alors représenté par un vecteur dense à valeurs réelles et de dimension $n$ fixe,

projeté dans un espace multidimensionnel continu : le plongement. Ces vecteurs sont préalablement appris sur de grandes quantités de données à l'aide d'approches basées sur des réseaux de neurones peu profonds. En plus du faible nombre de dimensions des vecteurs, ces algorithmes capture mieux la sémantique des mots que les vecteurs discrets. Autrement dit, deux mots proches sémantiquement auront des vecteurs semblables et seront proches dans l'espace de représentation.

Les travaux sur ces types de représentations vectorielles ont été initiés par Bengio et al. (2003). Les premières versions de plongements de mots ont été publiés en 2013 avec Word2vec, des plongements de mots en 300 dimensions, décliné en deux architectures : *Continuous Bag-of-words* (CBOW) (Mikolov et al., 2013a) et *Skip-gram* (Mikolov et al., 2013b). Suite à Word2vec et aux algorithmes CBOW et Skip-gram, d'autres méthodes de plongements de mots ont été publiés : GloVe (Pennington et al., 2014b) et fastText (Bojanowski et al., 2017). Nous présentons dans cette section ces algorithmes.

### 2.2.1 Plongements de mots Word2vec : algorithmes CBOW et Skip-gram

*Continuous bag-of-words* (CBOW) est la première architecture introduite par Mikolov et al. (2013a). Les représentations CBOW sont apprises à l'aide d'un réseau de neurones peu profond (un réseau à une seule couche cachée) ayant pour objectif de prédire un mot à partir de son contexte. Une illustration de l'architecture CBOW est donnée dans la figure 2.3. Pour créer un modèle CBOW, un mot cible et son contexte sont tout d'abord convertis en vecteurs *one-hot*. Un vecteur *one-hot* est une représentation où toutes les composantes sont 0 sauf une qui est 1. La couche d'entrée est la somme des vecteurs *one-hot* du contexte et la couche de sortie est le vecteur *one-hot* du mot cible. La couche cachée est la couche de plongements (*embeddings*) où chaque mot du vocabulaire est représenté par un vecteur de valeurs réelles. Enfin, le modèle est entraîné en comparant les plongements prédits et les plongements réels du mot cible et en corrigeant la représentation vectorielle du mot cible par rétro-propagation du gradient. Une fois le modèle CBOW entraîné, les plongements des mots se trouvent dans la couche cachée du réseau neuronal.

L'architecture *Skip-gram* a été publié par Mikolov et al. (2013b) la même année. À l'inverse de *CBOW*, l'architecture *Skip-gram* tente de prédire pour un mot donné le contexte dont il est issu. La couche d'entrée de ce réseau neuronal est alors un vecteur ne contenant qu'un seul mot, projeté dans la couche cachée puis dans la couche de sortie. De la même

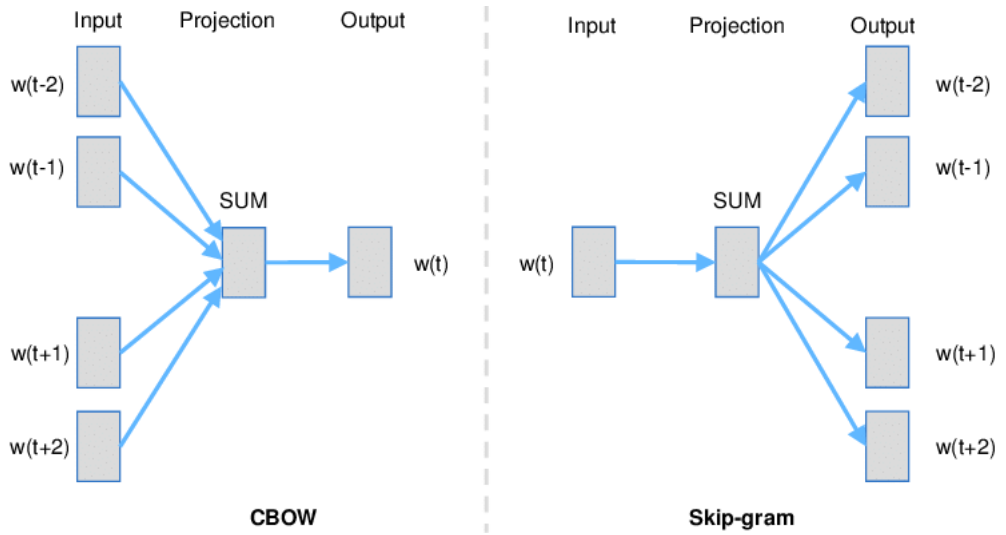


FIGURE 2.3 : Illustration des algorithmes *Continuous bag-of-words* (CBOW) et *Skip-gram*. Figure issue de l'article de Landthaler et al. (2017).

façon que dans l'architecture *CBOW*, le modèle est entraîné en comparant les plongements prédits et les plongements réels de chaque mot du contexte, puis en corrigeant les représentations vectorielles de ces mots par rétro-propagation du gradient. Cependant, dans un vocabulaire de grande taille, il y a un grand nombre de mots qui ne font pas partie du contexte d'un mot donné. Pour chaque paire positive `(mot cible, mot du contexte)`, le modèle doit générer de nombreuses paires négatives `(mot cible, mots du vocabulaire qui ne sont pas dans le contexte)` pour l'entraînement. Calculer le gradient et mettre à jour les poids pour chaque paire négative peut être très coûteux en termes de temps de calcul, surtout avec un vocabulaire volumineux.

Pour résoudre ce problème d'optimisation, Mikolov et al. (2013b) ont introduit, en même temps que *Skip-gram*, la notion d'échantillonnage négatif (*negative sampling*). Au lieu de générer toutes les paires négatives, un échantillonnage stochastique est appliqué pour ne générer qu'un sous-ensemble de paires négatives pour chaque exemple positif. Le nombre d'échantillons négatifs est généralement beaucoup plus petit que le nombre de mots dans le vocabulaire, ce qui accélère considérablement le processus d'entraînement du modèle.

### 2.2.2 Plongements de mots GloVe

Le modèle GloVe (Pennington et al., 2014a) (abréviation de *Global Vectors for Word Representation*) combine les avantages de deux méthodes pour la représentation de mots : (i) les méthodes basées sur la cooccurrence statistiques des mots (c'est-à-dire les matrice de cooccurrences, comme présenté dans la section 2.1) et (ii) les méthodes de plongements de mots comme Word2vec. Pour entraîner un modèle GloVe, une matrice de cooccurrences est construite à partir du corpus de textes. Comme présenté dans la section 2.1, cette matrice mesure combien de fois chaque paire de mots cooccurrent dans une fenêtre de contexte donnée. À partir de la matrice de cooccurrences, les probabilités de cooccurrence des mots sont calculées. Ces probabilités représentent la probabilité conditionnelle qu'un mot $j$ cooccure avec un mot $i$. Le modèle est ensuite entraîné par rétro-propagation du gradient en minimisant la différence entre les produits scalaires des vecteurs de mots et les logarithmes des probabilités de cooccurrence. Une fois l'entraînement terminé, les vecteurs de mots finaux sont obtenus.

### 2.2.3 Plongements de mots fastText

Les modèles Word2vec et GloVe ne tiennent pas compte de la structure interne des mots (préfixes, suffixes, racines et mots composés), chaque mot étant encodé par un vecteur distinct, sans partage de paramètres. Ainsi, les mots *dermatologie* et *épiderme* partageant une racine commune *derm-* auront des similitudes liées à leur contexte et non aux sous-mots qu'ils partagent. Pour pallier ce problème, fastText (Bojanowski et al., 2017) représente chaque mot à partir d'une combinaison de sous-mots, en les sommant. Intuitivement, il devient donc possible de construire une représentation d'un grand ensemble de mots partageant une même morphologie, même lorsque ces derniers n'apparaissent pas dans le corpus d'apprentissage. Dans son fonctionnement, l'architecture fastText reprend le principe des modèles CBOW et Skip-gram, à la seule différence que chaque mot est décomposé en sous-mots qui ont chacun leur représentation vectorielle. Afin de distinguer les préfixes et les suffixes d'autres chaînes de caractères, les caractères $<$ et $>$ sont ajoutés au début et à la fin des mots avant de créer les sous-mots à partir des mots. La représentation vectorielle finale d'un mot est obtenue en sommant les vecteurs de tous les sous-mots qui le composent.

L'apprentissage de plongements de sous-mots permet également de pouvoir représenter les mots hors vocabulaire. Pour créer la représentation d'un mot hors du vocabulaire du

corpus, il suffit donc de sommer les vecteurs des sous-mots qui composent ce mot pour obtenir son plongement.

## 2.3 Représentations continues et contextualisées

Les plongements de mots statiques présentés fournissent pour chaque mot une seule représentation vectorielle, quel que soit le contexte dans lequel le mot apparaît. Dans le cas des homographes, les représentations vectorielles seront moins fiables car ces mots ont pu être vus dans des contextes totalement différents. En effet, certains mots sont polysémiques et ne portent pas le même sens selon le contexte dans lequel ils sont vus. Par exemple, dans le domaine général, le mot *iris* peut se référer à une fleur. Alors que dans le domaine médical, le mot *iris* se réfère à l'iris de l'œil. La polysémie peut aussi être présente au sein d'un même domaine de spécialité. Par exemple, dans le domaine médical, le mot *ventricule* peut à la fois faire référence à un *ventricule cardiaque* ou à un *ventricule cérébral*. Avec l'introduction de deux nouvelles architectures : ELMo (Peters et al., 2018) et *Transformer* (Vaswani et al., 2017), un changement de paradigme s'est opéré, passant de plongements statiques, où chaque mot se voit attribuer un seul vecteur, à des plongements contextualisés, où le même mot peut avoir différentes représentations selon son contexte.

Nous décrivons dans cette section les deux principales architectures responsables de ce changement de paradigme : le modèle ELMo (Peters et al., 2018) et l'architecture *Transformer* (Vaswani et al., 2017). Nous présentons ensuite plusieurs modèles de plongements contextualisés ayant comme architecture commune le *Transformer*, notamment BERT (Devlin et al., 2019) et GPT (Radford et al., 2018) qui ont ouvert la voie aux modèles de langue pré-entraînés tels que nous les connaissons aujourd'hui. Enfin, nous introduisons les méthodes permettant d'adapter les modèles de langue pré-entraînés à des domaines de spécialité où les données sont disponibles en plus petite quantité. Nous illustrons avec les domaines biomédical et clinique en mettant en avant les modèles existants dans ces domaines.

### 2.3.1 ELMo

ELMo (Peters et al., 2018) (*Embeddings from Language Models*) est une architecture neuronale conçue dans le but de produire des représentations de mots contextualisées.

Plus précisément, le modèle utilise une architecture basée sur des couches récurrentes bidirectionnelles (*Bidirectional Long Short-term Memory* ou *BiLSTM*) qui est d'abord entraînée sur une tâche de modélisation de langage (c'est-à-dire la prédiction du mot suivant dans une phrase). Ensuite, les représentations internes du modèle sont combinées pour former des représentations de mots contextualisées.

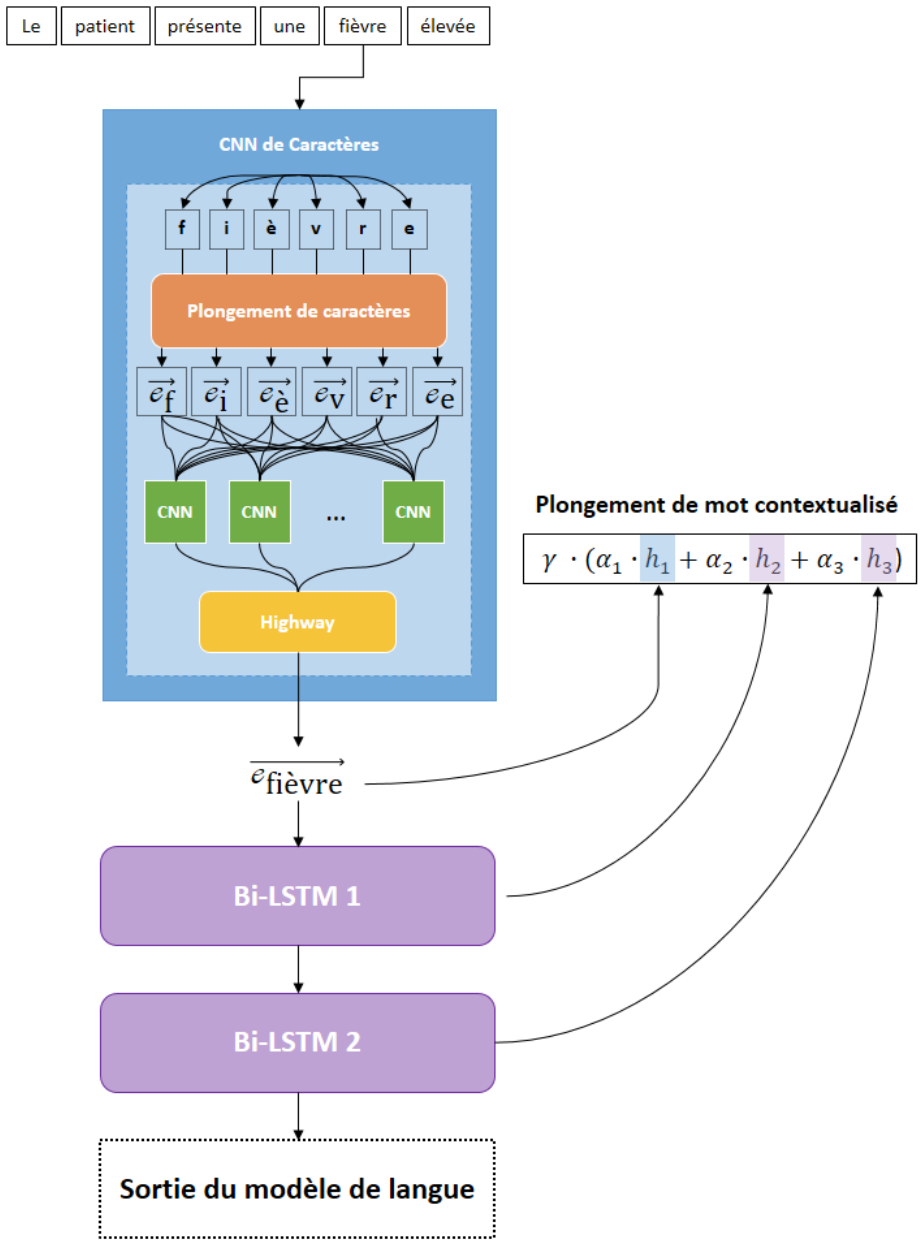


FIGURE 2.4 : Architecture du modèle ELMo et obtention des plongements de mots contextualisés. Figure inspirée de la thèse de El Boukkouri (2021)

L'architecture d'ELMo est composé de trois modules : un réseau de neurones à convolution (*convolutional neural network* ou *CNN*) (LeCun et al., 2015) pour les caractères et deux Bi-LSTMs (Hochreiter and Schmidhuber, 1997) (figure 2.4). Le CNN pour les caractères convertit les mots en entrée en représentations statiques avant qu'ils ne puissent être contextualisés à l'aide des Bi-LSTM en aval. Le processus exact implique de convertir chaque mot en une séquence de caractères, créer un plongement pour chaque caractère et d'utiliser un ensemble de couches CNN pour effectuer des convolutions sur cette séquence. Les sorties du CNN sont ensuite combinées en un vecteur final qui peut être projeté sur une taille souhaitée. Ensuite, les deux couches Bi-LSTM se basent sur ce plongement statique pour produire une représentation contextualisée qui sert de fondation pour la tâche de pré-entraînement de modélisation de langage.

Chaque BiLSTM est composé de deux couches : (1) une couche en avant (*forward layer*) et (2) une couche en arrière (*backward layer*). Pour une séquence de $N$ mots $(t_1, t_2, ...t_N)$, la couche en avant traite la séquence de mots dans le sens de la lecture (de gauche à droite) et calcule la probabilité de la séquence en modélisant la probabilité du mot $t_k$ étant donné le contexte précédent $(t_1, t_2...t_{k-1})$ :

$$p(t_1, t_2, ...t_N) = \prod_{k=1}^{N} p(t_k|t_1, t_2, ..., t_{k-1}) \tag{2.1}$$

À l'inverse, la couche en arrière traite la séquence de mots dans le sens inverse de lecture (de droite à gauche) et calcule la probabilité de la séquence en modélisant la probabilité du mot $t_k$ étant donné le contexte suivant $(t_{k+1}, t_{k+2}...t_N)$ :

$$p(t_1, t_2, ...t_N) = \prod_{k=1}^{N} p(t_k|t_{k+1}, t_{k+2}, ..., t_N) \tag{2.2}$$

Finalement, une fois le pré-entraînement terminé, la représentation finale d'un mot $w$ peut être obtenue en combinant les représentations des différentes couches en un vecteur final :

$$\text{ELMo}(w) = \gamma \cdot (\alpha_1 \cdot h_1 + \alpha_2 \cdot h_2 + \alpha_3 \cdot h_3) \tag{2.3}$$

où $\gamma$, $\alpha_1$, $\alpha_2$ et $\alpha_3$ sont des poids ajustables et $h_1$, $h_2$ et $h_3$ sont respectivement les représentations en sortie du CNN de caractères, du premier et du deuxième BiLSTM.

### 2.3.2 Architecture Transformer

Les *Transformers* (Vaswani et al., 2017) sont les principaux composants des modèles de langue pré-entraînés récents comme BERT (Devlin et al., 2019) ou GPT (Radford et al., 2018). L'architecture *Transformer* (Vaswani et al., 2017) est relativement complexe et a été développée à l'origine pour des tâches de séquence à séquence comme la traduction automatique. Celle-ci est composée d'un encodeur et d'un décodeur, comme présenté dans la figure 2.5. L'encodeur transforme une séquence de mots en entrée $(x_1, ..., x_n)$ en une séquence de représentations continues $z = (z_1, ..., z_n)$. En utilisant $z$, le décodeur génère ensuite un par un les mots d'une séquence de sortie $(y_1, ..., y_m)$. À chaque étape, le modèle est auto-régressif, c'est-à-dire qu'il utilise les mots précédemment générés comme entrée additionnelle lors de la génération du mot suivant. Même si cette architecture a pu être adaptée par la suite à différents cas d'application, par exemple l'utilisation de l'encodeur seul dans BERT ou l'utilisation du décodeur seul dans GPT, chaque module a principalement été réutilisé tel que défini par les auteurs.

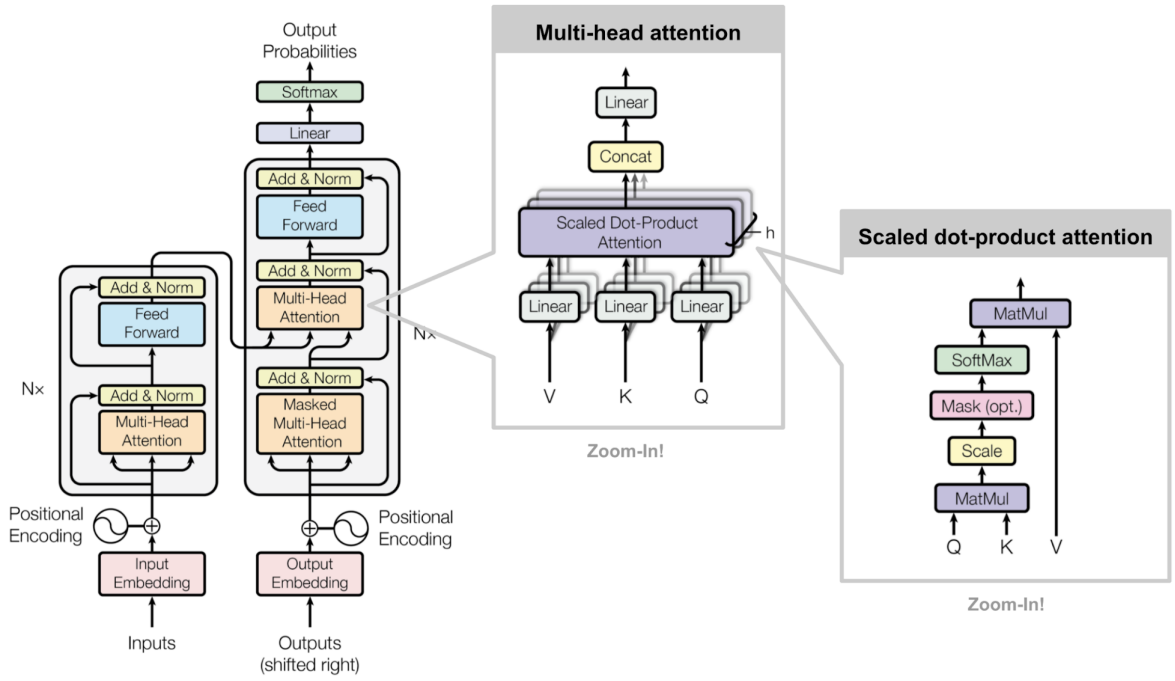


FIGURE 2.5 : Architecture Transformer. Figure originale extraite de l'article de Vaswani et al. (2017).

L'encodeur est composé d'un bloc d'auto-attention multi-têtes (*Multi-Head Attention*) suivi d'un bloc de réseau à propagation avant (*Feed-Forward*). Le décodeur est également

composé d'un bloc d'auto-attention multi-têtes suivi d'un bloc de réseau à propagation avant, avec en plus un bloc d'attention *Encoder-Decoder* permettant de faire le lien entre la séquence encodée en entrée, et la séquence de sortie qui est décodée.

Le mécanisme d'auto-attention permet de mesurer le lien entre un mot et tous les autres mots d'une même séquence. De ce fait, chaque mot de la séquence d'entrée peut être traité de manière séparée. L'auto-attention est ici multi-têtes, cela signifie qu'un bloc d'auto-attention est composé de plusieurs têtes d'attention, permettant ainsi d'exécuter plusieurs fois le mécanisme d'attention en parallèle. Ce mécanisme d'auto-attention repose sur trois paramètres appelés requête (Q pour *Query*), clé (K pour *Key*) et valeur (V pour *Value*). Chacun de ses paramètres est représenté par un vecteur pour chaque mot de la séquence. À partir de la séquence de mots en entrée et de ces trois paramètres, un score d'attention est calculé pour chaque mot en entrée par une somme pondérée des valeurs $v_{t1 \leq t \leq n}$ en utilisant des poids qui reflètent la similarité entre la requête $q$ du mot actuel et les clés $k_{t1 \leq t \leq n}$ de tous les autres mots de la séquence. Les vecteurs de requête, clé et valeur sont appris en tant que projections des représentations d'entrée, et le calcul global de l'attention est répété par plusieurs têtes, chacune ayant des paramètres $Q$, $K$ et $V$ différents. Les sorties de toutes les têtes sont concaténées et projetées pour produire la séquence finale de représentations contextualisées.

$$\begin{aligned} MultiHead(Q,K,V) &= Concat(head_1, ..., head_h)W^O \\ \text{avec } head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \\ \text{et } Attention(Q,K,V) &= softmax\left(\frac{QK^t}{\sqrt{d_k}}\right)V \end{aligned} \tag{2.4}$$

où les matrices de paramètres $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$ et $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ sont respectivement les matrices pour les requêtes, les clés et les valeurs pour la tête d'attention $i$. $W^O \in \mathbb{R}^{hd_v \times d_{model}}$ est la matrice de projection finale du mécanisme d'attention. Toutes ces matrices sont apprises durant l'apprentissage. $d_k$ est la dimension des requêtes et des clés, $d_v$ est la dimension des valeurs et $d_{model}$ est la dimension des représentations vectorielles en sortie du modèle. Dans l'article présentant l'architecture *Transformer*, les auteurs ont utilisé $h = 8$ têtes d'attention et pour chaque tête $d_k = d_v = d_{model}/h = 64$.

Un autre composant est crucial au bon fonctionnement du *Transformer* pour la création de représentations contextualisées : l'encodage positionnel. Contrairement aux réseaux de neurones récurrents où l'ordre des mots est naturellement encodé par le proces-

sus de récurrence, la position des mots dans la séquence n'est pas prise en compte dans le calcul de l'attention du *Transformer*, c'est-à-dire qu'un même vecteur de mot en entrée dans différentes positions aura la même représentation. Une méthode d'encodage de la position des mots dans la séquence est donc nécessaire pour inclure cette information dans les vecteurs. Pour cela, les auteurs du *Transformer* ont proposé l'encodage positionnel (*positional encoding*) qui consiste à ajouter à chaque vecteur de mot en entrée (*input embedding*) une représentation vectorielle basée sur une sinusoïde de dimension $d$ (identique au nombre de dimensions du vecteur de mot en entrée) qui dépend de la position $p$ du mot dans la séquence. L'encodage positionnel est formulé comme ceci :

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right) \tag{2.5}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right) \tag{2.6}$$

où *pos* est la position et $i$ est la dimension.

## 2.4 Modèles de langue pré-entraînés

Depuis l'avènement des *Transformers* (Vaswani et al., 2017) et de l'auto-attention, beaucoup de modèles utilisant cette architecture ont émergé et s'imposent comme approches état de l'art dans la grande majorité des tâches de TAL. Les performances de ces modèles évoluent selon leur nombre de paramètres et la quantité de données sur lesquelles ils sont pré-entraînés. De ce fait, les modèles pré-entraînés sont de plus en plus grands au fil du temps, avec des modèles atteignant plusieurs milliards de paramètres aujourd'hui. L'entraînement de tels modèles s'effectue *via* des tâches auto-supervisées ou auto-régressives sur des données non étiquetées, permettant ainsi de se défaire des représentations de mots apprises au préalable puisque celles-ci sont apprises intrinsèquement dans le modèle.

Ces modèles s'appuient maintenant sur des algorithmes de segmentation des mots (*tokenizer*) entraînés sur le même corpus afin de créer un vocabulaire de sous-mots (*tokens*). La segmentation en tokens permet au modèle d'avoir une taille de vocabulaire raisonnable (généralement entre 16 000 et 52 000 tokens) tout en étant capable d'apprendre des représentations sémantiques. De plus, la segmentation en tokens permet au modèle de traiter les mots hors vocabulaire en les décomposant en tokens connus, en allant parfois jusqu'au niveau caractère si cela est nécessaire. Une fois ces modèles pré-entraînés, un

affinage (*fine-tuning*) du modèle peut être réalisé sur les tâches traditionnelles de TAL comme la reconnaissance d'entités nommées ou la classification de texte.

Dans cette section, nous présentons tout d'abord les *tokenizers* qui sont employés dans les modèles de langue pré-entraînés. Nous décrivons ensuite les apprentissages auto-supervisé et auto-régressif, et plus précisément, les tâches de ces paradigmes permettant le pré-entraînement des modèles de langue. Nous présentons ensuite plusieurs modèles de langue pré-entraînés fondateurs se basant sur l'architecture *Transformer*, que nous avons divisés en trois catégories : (i) les modèles encodeur, illustrés avec le modèle BERT et ses variantes ; (ii) les modèles décodeur avec le modèle GPT et ses successeurs et (iii) les modèles encodeur-décodeur, illustrés avec les modèles BART (Lewis et al., 2020) et T5 (Raffel et al., 2020).

### 2.4.1 Algorithmes de segmentation en sous-mots

Les algorithmes de segmentation en sous-mots (ou appelés *tokenizers*) utilisés dans les modèles de langue se basent sur le principe que les mots fréquents ne doivent pas être segmentés et, à l'inverse, les mots rares doivent être décomposés en sous-mots. Par exemple, dans le domaine médical, le terme « cardiopathie » est un mot rare décomposé en « cardio » et « pathie ». En tant que sous-mots, « cardio » et « pathie » apparaissent plus fréquemment que cardiopathie. La sémantique de « cardiopathie » est calculée par composition en agrégeant les sens de « cardio » et « pathie ». Cette modélisation est particulièrement pertinente pour les langues de spécialité qui font usage de racines gréco-latines.

Nous présentons les principaux algorithmes de segmentation en sous-mots : Byte-Pair Encoding (BPE) (Sennrich et al., 2016), WordPiece (Schuster and Nakajima, 2012), Unigram (Kudo, 2018) et SentencePiece (Kudo and Richardson, 2018). Tous ces algorithmes s'appuient sur un corpus de textes pour générer le vocabulaire de sous-mots. Généralement, ce corpus est le même que celui sur lequel le modèle de langue sera ensuite pré-entraîné.

**Byte-Pair Encoding (BPE)** L'algorithme de segmentation *Byte-Pair Encoding* (BPE) (Sennrich et al., 2016) est initialement inspiré d'un algorithme de compression introduit par Gage (1994). BPE est un algorithme ascendant : il commence avec une liste de caractères uniques et construit son vocabulaire en examinant la fréquence des séquences de caractères.

L'algorithme BPE commence par une étape de pré-segmentation pour diviser le corpus

en mots. Cette pré-segmentation peut être réalisée simplement avec une segmentation par espaces ou avec des outils plus avancés utilisant des systèmes à base de règles, comme Moses (Koehn et al., 2007) ou Spacy. Après cette étape de pré-segmentation, une liste des mots présents dans le corpus est créée où chaque mot est associé à sa fréquence d'apparition dans le corpus. Ensuite, un vocabulaire $\mathcal{V}$ est initialisé avec tous les caractères présents dans la liste de mots. La suite de l'algorithme repose sur deux étapes qui sont répétées jusqu'à ce que la taille du vocabulaire souhaitée soit atteinte. À noter que la taille du vocabulaire souhaitée $\mathcal{M}$ est un des hyperparamètres à définir avant de lancer l'algorithme.

Jusqu'à ce que la taille du vocabulaire souhaitée soit atteinte (c'est-à-dire lorsque $|\mathcal{V}| = M$) :

1. La fréquence de chaque paire de tokens du vocabulaire $\mathcal{V}$ observée dans le corpus est calculée.

2. La paire de tokens la plus fréquence est fusionnée puis ajoutée en tant que token dans le vocabulaire $\mathcal{V}$.

La taille finale du vocabulaire correspond donc au nombre de caractères auquel est ajouté le nombre de fusions de paires de tokens.

**WordPiece** L'algorithme de segmentation WordPiece (Schuster and Nakajima, 2012) est similaire à l'algorithme BPE. À l'instar de BPE, WordPiece initialise le vocabulaire avec chaque caractère du corpus, puis ajoute progressivement les fusions jusqu'à atteindre la taille de vocabulaire souhaitée. La différence avec BPE réside dans la manière de choisir la fusion qui sera ajoutée au vocabulaire. Contrairement à BPE, WordPiece ne choisit pas la paire de tokens la plus fréquente mais la paire qui, une fois ajoutée au vocabulaire, maximise la log-vraisemblance sur l'ensemble du corpus.

**Unigram** Unigram (Kudo, 2018) est un algorithme de segmentation en sous-mots capable de produire, pour un même mot, plusieurs segmentations avec des probabilités. Contrairement à BPE ou WordPiece, Unigram est un algorithme descendant : le vocabulaire est initialisé avec un grand nombre de tokens qui est ensuite réduit progressivement jusqu'à obtenir la taille de vocabulaire souhaitée. Unigram n'est pas utilisé directement dans les modèles type *Transformer* mais en conjonction avec l'algorithme SentencePiece que nous présentons dans la section suivante.

Dans Unigram, on part de l'hypothèse que chaque sous-mot apparaît indépendamment des autres. Pour un mot $\mathbf{x}$ et une segmentation de ce mot en sous-mots $s(\mathbf{x}) = (x_1, ..., x_L)$, la probabilité de la segmentation $s(\mathbf{x})$ est formulée comme le produit des probabilités d'occurrences des sous-mots $(x_1, ..., x_L)$ dans le corpus :

$$P(s(\mathbf{x})) = \prod_{i=1}^{L} p(x_i) \tag{2.7}$$

En pratique, $p(x_i)$ est estimé en comptant les occurrences de $x_i$ dans le corpus par rapport au nombre total de sous-mots dans le corpus. Pour un mot $\mathbf{x}$ et un ensemble de segmentations candidates $\mathcal{S}(\mathbf{x})$, la segmentation la plus probable $s(\mathbf{x})^*$ est obtenu comme suit avec l'algorithme Viterbi (Viterbi, 1967) :

$$s(\mathbf{x})^* = \arg\max_{s(\mathbf{x})\in\mathcal{S}(\mathbf{x})} P(s(\mathbf{x})) \tag{2.8}$$

À partir de ces probabilités d'occurrences des segmentations, la log-vraisemblance $\mathcal{L}$ est estimée sur l'ensemble du corpus :

$$\mathcal{L} = \sum_{\mathbf{x}\in C} \log\left(\sum_{s(\mathbf{x})\in\mathcal{S}(\mathbf{x})} P(s(\mathbf{x}))\right) \tag{2.9}$$

où $\mathbf{x}$ est un mot du corpus C, $s(\mathbf{x})$ est une segmentation candidate du mot $\mathbf{x}$ et $\mathcal{S}(\mathbf{x})$ sont toutes les segmentations candidates du mot $\mathbf{x}$.

La première étape de l'algorithme Unigram est la pré-segmentation du corpus. Le vocabulaire $\mathcal{V}$ est ensuite initialisé avec tous les caractères uniques ainsi que les séquences de caractères les plus fréquentes. Les étapes suivantes sont ensuite répétées jusqu'à obtenir la taille de vocabulaire souhaitée (c'est-à-dire lorsque $|\mathcal{V}| = M$) :

1. La log-vraisemblance $\mathcal{L}$ est calculée sur le corpus entier à partir du vocabulaire courant.

2. Pour chaque token $x_i$ du vocabulaire, on calcule une fonction de perte ($loss_i$) qui estime à quel point $\mathcal{L}$ augmenterait si le token $x_i$ était retiré du vocabulaire actuel.

3. Tous les tokens du vocabulaire sont ensuite triés par valeur de $loss_i$ et les $\eta$ % des tokens qui ont les plus petites $loss_i$ sont retirés du vocabulaire. $\eta$ est un hyperparamètre généralement défini à 20 %. Les tokens retirés du vocabulaire ont le

moins d'impact sur la log-vraisemblance $\mathcal{L}$ et sont par conséquent considérés comme moins nécessaires. À noter que les caractères uniques ne peuvent pas être retirés du vocabulaire afin de pouvoir toujours représenter les mots hors vocabulaire.

Comme Unigram n'effectue pas fusions de paires comme BPE et WordPiece, une fois l'entraînement de l'algorithme terminé, un mot peut être segmenté de plusieurs façons possibles. Par exemple, avec le vocabulaire Unigram suivant `["a", "e", "t", "r", "b", "art", "ère", "ar"]`, le mot « artère » pourrait être segmenté comme `["ar", "t", "ère"]`, `["art", "ère"]` ou `["a", "r", "t", "ère"]`. En plus de sauvegarder le vocabulaire, Unigram sauvegarde également les probabilités calculées pour token dans le corpus, permettant ainsi de calculer après l'entraînement du modèle la probabilité de chaque segmentation possible. En pratique, c'est toujours la segmentation la plus probable qui est retenue.

**SentencePiece** Tous les algorithmes de segmentation présentés précédemment partent du principe que les mots du texte sont séparés par des espaces et reposent tous, par conséquent, sur une première étape de pré-segmentation exploitant les séparateurs de mots. Cependant, pour les langues avec un système d'écriture continuen, des outils de pré-segmentation propre à la langue sont nécessaires. L'algorithme SentencePiece (Kudo and Richardson, 2018) propose une solution qui n'intègre pas d'étape de pré-segmentation et traite la séquence au niveau phrase avec les algorithmes BPE, WordPiece ou Unigram. Pour les langues utilisant des espaces pour séparer les mots, les espaces sont remplacés par le caractère spécial « ▁ » qui est inclus dans le vocabulaire. SentencePiece intègre également dans son algorithme le processus de normalisation au niveau des caractères NFKC. Ce processus élimine les ambiguïtés Unicode entre les caractères ayant les mêmes glyphes (c'est-à-dire qui paraissent identiques) dans différentes langues mais qui sont associés à différents identifiants Unicode.

### 2.4.2 Apprentissages auto-supervisé et auto-régressif

Les apprentissages auto-supervisé et auto-régressif des modèles de langue sont à la croisée entre l'apprentissage supervisé et non supervisé. Ces paradigmes d'apprentissage, à partir desquelles le modèle est pré-entraîné, tirent parti de l'information intrinsèque contenue dans les données non étiquetées pour créer des tâches de prédiction. Il existe de multiples objectifs de pré-entraînement d'un modèle, nous présentons ici les deux principaux ob-

jectifs : le *Causal Language Modeling* (ou CLM) pour l'apprentissage auto-régressif et le *Masked Language Modeling* (ou MLM) pour l'apprentissage auto-supervisé.

Le *Causal Language Modeling* est l'objectif classique d'une modélisation unidirectionnelle. Cet objectif de pré-entraînement est une tâche de prédiction du token suivant étant donné le contexte précédent, comme illustré dans la figure 2.6.

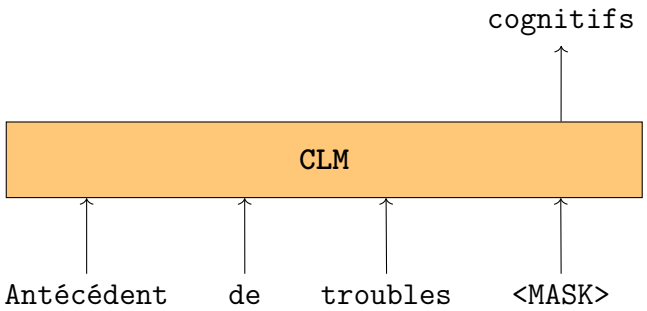


FIGURE 2.6 : Illustration du *Causal Language Modeling* (pour des raisons de simplicité, un *token* correspond à un mot dans cette illustration).

Autrement dit, on cherche à estimer la distribution de probabilité d'une séquence de $N$ tokens, soit de gauche à droite :

$$p(t_1, t_2, ..., t_N) = \prod_{k=1}^{N} p(t_k | t_1, t_2, ..., t_{k-1}) \tag{2.10}$$

Soit de droite à gauche :

$$p(t_1, t_2, ...t_N) = \prod_{k=1}^{N} p(t_k | t_{k+1}, t_{k+2}, ..., t_N) \tag{2.11}$$

On retrouve cette tâche dans les réseaux neuronaux récurrents comme les BiLSTM du modèle ELMo ou les modèles GPT. Le *Causal Language Modeling* peut être vu comme un cas particulier du *Masked Language Modeling*.

Le *Masked Language Modeling* est un objectif de modélisation bidirectionnel qui consiste à remplacer un pourcentage des tokens de la séquence par des tokens spécifiques `<MASK>`. Le modèle a ensuite comme tâche de prédire les *tokens* masqués à partir du contexte restant de la séquence, comme illustré en figure 2.7.

Le *Masked Language Modeling* a été introduit avec le modèle BERT (Devlin et al., 2019). Pour une séquence de *tokens* donnée $(t_1, ..., t_N)$ et une fonction de masquage $\phi$, on

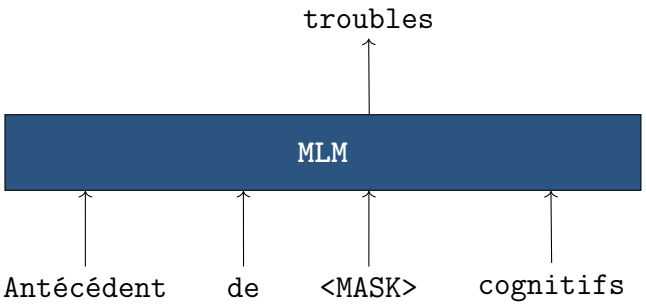


FIGURE 2.7 : Illustration du *Masked Language Modeling* (pour des raisons de simplicité, un *token* correspond à un mot dans cette illustration).

cherche à estimer la probabilité :

$$p\left((t_1, ..., t_N)|\phi((t_1, ..., t_N))\right) \tag{2.12}$$

où $\phi$ retourne une séquence possiblement altérée avec des *tokens* spéciaux.

En pratique, 15 % des *tokens* sont échantillonnés pour chaque séquence de *tokens*. Parmi ces 15 %, 80 % sont remplacés par le *token* spécial `<MASK>`, 10 % sont remplacés par un *token* aléatoire du vocabulaire et 10 % restent inchangés.

Le *Masked Language Modeling* se décline en plusieurs stratégies, le masquage des *tokens* (c'est-à-dire des sous-mots) ou le masquage des mots entiers (appelé *Whole Word Masking*) où tous les *tokens* qui composent un mot sont masqués.

Suite à l'introduction du *Masked Language Modeling*, plusieurs variantes ont vu le jour, créant ainsi une famille d'approches de modélisation de la langue : le *Denoising Language Modeling* (que l'on pourrait traduire simplement par modélisation de langage par débruitage). Le *Denoising Language Modeling* consiste à entraîner un modèle à débruiter un texte intentionnellement altéré pour le ramener à sa forme originale. Le *Denoising Language Modeling* est formulé comme le *Masked Language Modeling* (voir équation 2.12), seule la fonction d'altération du texte $\phi$ change. Dans le *Masked Language Modeling*, l'altération du texte se fait sous la forme d'un masquage mais il existe d'autres procédés pour altérer le texte. Par exemple, dans le modèle BART (Lewis et al., 2020), l'altération du texte est effectuée en mélangeant aléatoirement les phrases de la séquence et en masquant des séquences de tokens.

### 2.4.3 Modèle Encodeur : BERT et ses variantes

Les modèles Encodeur sont composés uniquement d'encodeurs *Transformer*. Ils sont conçus pour capturer la représentation contextuelle des mots dans un texte en prenant en compte le contexte à gauche et à droite de chaque mot. Ces modèles sont utilisés dans toutes les tâches de TAL telles que les tâches de classification, de compréhension de texte ou d'extraction d'entités.

Parmi les modèles de langue fondateurs s'appuyant uniquement sur des encodeurs, BERT (Devlin et al., 2019), pour *Bidirectional Encoder Representations from Transformers*, est le plus courant. Le modèle BERT est décomposé en deux étapes : le pré-entraînement (*pre-training*) et l'affinage (*fine-tuning*). Le pré-entraînement de BERT est auto-supervisé sur des données non étiquetées, entraîné avec les tâches de *Masked Language Modeling* (introduit dans la section 2.4.2) et de prédiction de la phrase suivante (*Next-Sentence Prediction*), comme illustré dans la partie *Pre-training* de la figure 2.8. Pour son pré-entraînement, BERT prend en entrée deux phrases distinctes A et B qui sont jointes par un *token* spécial de séparation `<SEP>`. L'objectif de pré-entraînement de prédiction de la phrase suivante consiste à prédire la phrase B à partir de la phrase A. Un *token* spécial `<CLS>` a été introduit au début de la séquence pour représenter la relation entre les deux phrases, comme illustré dans la partie *Pre-training* de la figure 2.8 avec la sortie *NSP*. Pour l'étape d'affinage, les paramètres du modèle sont initialisés avec les paramètres de l'étape de pré-entraînement, puis tous les paramètres du modèle sont mis à jour par apprentissage supervisé sur la tâche cible avec des données étiquetées. Chaque tâche cible a son modèle affiné séparé, même si les modèles de toutes ces tâches cibles sont initialisées avec les mêmes paramètres du modèle pré-entraîné. Un schéma de l'étape d'affinage pour la tâche de questions-réponses est donné dans la figure 2.8.

Le modèle BERT utilise l'algorithme WordPiece (introduit en section 2.4.1) avec une taille de vocabulaire de 30 000. Les représentations WordPiece sont ensuite sommées avec deux autres représentations : (i) le *segment embedding* qui indique à quelle phrase (A ou B) appartient chaque *token* et (ii) le *position embedding* qui indique la position de chaque *token* dans la séquence. Un exemple d'entrée du modèle est donné dans la figure 2.9.

À son origine, BERT est disponible en deux tailles : $\text{BERT}_{\texttt{BASE}}$ et $\text{BERT}_{\texttt{LARGE}}$. $\text{BERT}_{\texttt{BASE}}$ est composé de 12 couches d'encodeur *Transformer*, 12 têtes d'attention, une taille de sortie de 768, pour un total de 110 millions de paramètres. Pour ce qui est de $\text{BERT}_{\texttt{LARGE}}$, il est composé de 24 couches, 16 têtes d'attention, une taille de sortie de 1 024 et 340 millions de paramètres. Les deux modèles ont été entraînés sur la même quantité de

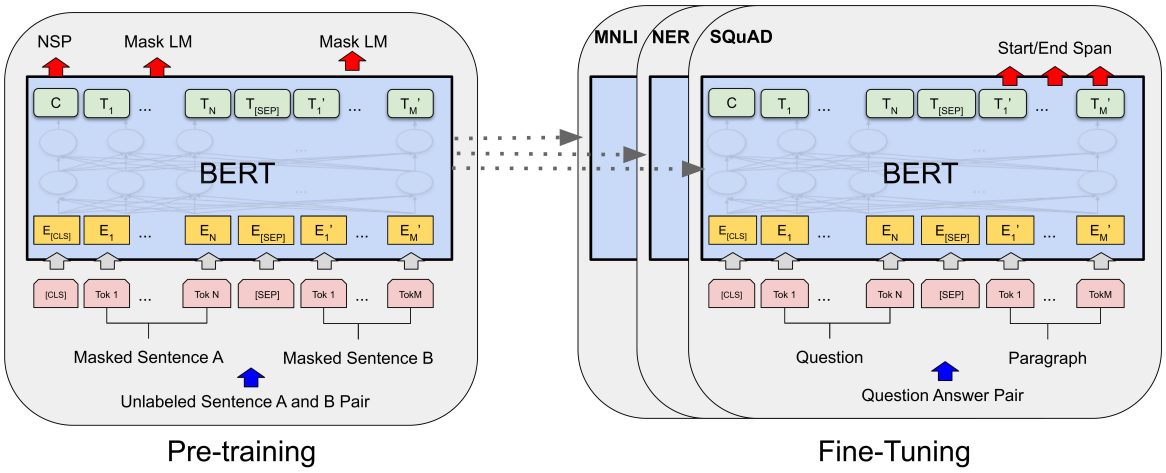


FIGURE 2.8 : Illustration du fonctionnement du modèle BERT. La partie *Pre-training* correspond au pré-entraînement auto-supervisé du modèle avec les tâches de *Masked Language Model* et de *Next Sentence Prediction*. La partie *Fine-Tuning* correspond à l'affinage du modèle sur une tâche cible, ici illustré avec la tâche de questions-réponses (*Question Answering* - *QA*). Figure extraite de l'article de Devlin et al. (2019).

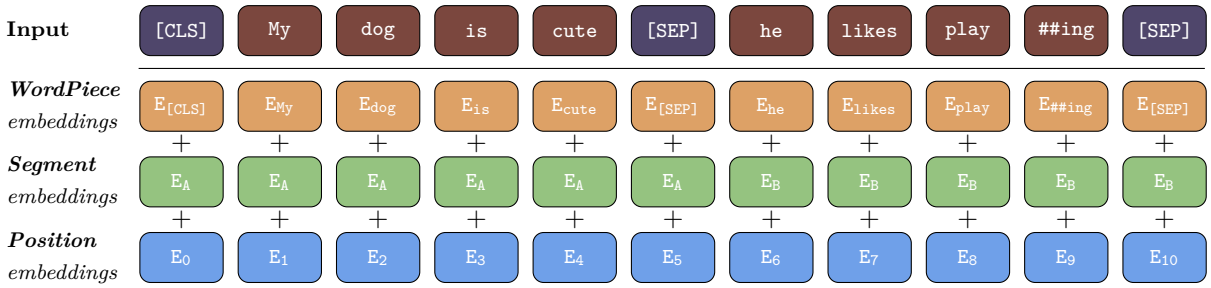


FIGURE 2.9 : Aperçu de l'entrée de BERT. Les deux séquences en entrée sont jointes en utilisant les *tokens* spéciaux `<CLS>` et `<SEP>`. Les *tokens* sont ensuite représentés avec l'algorithme WordPiece puis sommés avec les plongements de phrases et de positions.

données : 3,2 milliards de mots issus des corpus anglais Wikipedia (2,5 milliards de mots) et BooksCorpus (Zhu et al., 2015) (800 millions de mots). Le modèle $\text{BERT}_{\texttt{BASE}}$ a été conçu pour avoir la même taille que le modèle GPT, que nous présentons dans la section 2.4.4, afin de pouvoir comparer les deux modèles.

Suite au modèle BERT, plusieurs déclinaisons du modèle ont été publiées. Par exemple, le modèle RoBERTa (Liu et al., 2019) est une version améliorée de BERT avec quatre principales modifications :

1. RoBERTa est pré-entraîné uniquement sur une tâche de *Masked Language Modeling* dynamique où le masquage est réalisé à la volée lorsqu'une phrase est donnée au

modèle. En outre, une même phrase vu dix fois lors de l'apprentissage sera masquée de dix façons, contrairement à BERT où le masquage de la phrase est unique pour toutes les itérations d'apprentissage.

2. Liu et al. (2019) ont montré que la tâche de prédiction de la phrase suivante (*Next Sentence Prediction - NSP*) n'était pas utile pour la majorité des tâches cibles lors de l'affinage du modèle. Cette tâche de pré-entraînement a donc été retirée dans RoBERTa afin de pouvoir pré-entraîner le modèle sur des séquences continues de 512 *tokens*.

3. L'algorithme BPE avec une taille de vocabulaire de 50 000 est utilisé au lieu de WordPiece avec 30 000 de vocabulaire dans le modèle BERT original.

4. RoBERTa est entraîné avec une taille de *mini-batch* plus élevé, permettant un apprentissage plus court mais nécessitant des ressources de calcul en parallèle plus conséquentes.

Concernant les caractéristiques de l'architecture, RoBERTa reprend les mêmes caractéristiques que $\text{BERT}_{\texttt{LARGE}}$, à savoir, 24 couches, 16 têtes d'attention, une taille de sortie de 1 024 et 355 millions de paramètres au total. Enfin, RoBERTa est entraîné sur un jeu de données 10 fois plus grand que BERT (160 GB de texte pour RoBERTa contre 16 GB pour BERT).

Les besoins computationnels et de mémoire de l'auto-attention du *Transformer* augmentent quadratiquement avec la longueur de la séquence d'entrée, ce qui rend très coûteux le traitement de longues séquences. Par conséquent, BERT et ses variantes se limitent à des séquences de 512 *tokens* en entrée. Ces limites peuvent avoir des incidences pour le traitement des longs documents dans certaines tâches cibles.

Pour réduire cette complexité, plusieurs travaux sur les transformers éparses (*sparse transformers*) ont proposé des mécanismes d'attention alternatifs (Child et al., 2019; Zaheer et al., 2020; Kitaev et al., 2020; Wang et al., 2020; Tay et al., 2020; Beltagy et al., 2020; Choromanski et al., 2020; Katharopoulos et al., 2020). L'une des alternatives proposées est le Longformer (Beltagy et al., 2020), un modèle permettant de traiter des séquences allant jusqu'à 4 096 *tokens*. Afin de pouvoir traiter des séquences d'une telle taille, l'architecture *Transformer* a été modifié avec une opération d'auto-attention qui évolue linéairement avec la longueur de la séquence.

Le mécanisme d'auto-attention dans Longformer se compose à la fois d'un contexte local et d'un contexte global. Pour le contexte local, tous les *tokens* voient uniquement les

*tokens* dans une fenêtre de contexte $w$ définie, ce qui signifie que chaque *token* voit ses $\frac{w}{2}$ *tokens* précédents et ses $\frac{w}{2}$ *tokens* suivants. Pour le contexte global, certains *tokens* de la séquence sont sélectionnés pour accéder à tous les autres *tokens* de la séquence. À noter que cette opération d'attention est symétrique, un *token* avec une attention globale peut accéder à tous les *tokens* de la séquence et tous les *tokens* de la séquence ont accès à ce *token*. Les *tokens* ayant accès à la fois aux contextes local et global sont projetés à l'aide de matrices de requête, de clé et de valeur différentes : $(Q_s, K_s, V_s)$ pour l'attention locale et $(Q_g, K_g, V_g)$ pour l'attention globale.

Le modèle Longformer est ensuite pré-entraîné avec le *Masked Language Modeling* avec les paramètres de RoBERTa afin que le modèle converge plus rapidement et que cela soit moins coûteux en ressources computationnelles.

### 2.4.4 Modèle Décodeur : GPT et ses successeurs

Les modèles Décodeur sont conçus uniquement avec des décodeurs *Transformer*. Ils prennent en entrée un contexte initial (souvent appelé amorce ou *prompt*) et génèrent ensuite du texte séquentiellement en fonction de ce contexte. Ces modèles sont généralement utilisés pour la génération de texte créatif, la complétion de texte, les réponses automatiques, etc.

GPT (Radford et al., 2018) fut le premier modèle de langue pré-entraîné dont l'architecture était composée uniquement de décodeurs *Transformer*. À l'instar de BERT, le modèle GPT est décomposé en deux étapes : le pré-entraînement et l'affinage. Le pré-entraînement de GPT est auto-régressif sur des données non étiquetées, entraîné sur la tâche de *Causal Language Modeling* (présenté dans la section 2.4.2) avec des séquences continues de 512 *tokens* en entrée. L'architecture de GPT est composée de 12 couches de décodeurs Transformer, 12 têtes d'attention, une taille de sortie de 768. L'algorithme BPE (présenté dans la section 2.4.1) avec une taille de vocabulaire fixé 40 000. Le modèle a été entraîné sur le corpus BooksCorpus (Zhu et al., 2015).

De nouvelles versions de GPT ont suivi, GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020) et GPT-4 (OpenAI, 2023), avec à chaque nouvelle version une augmentation du nombre de paramètres et de la quantité de données de pré-entraînement. GPT-3, le dernier modèle pour lequel nous avons les détails sur le nombre de paramètres et la quantité de données d'apprentissage, possède 175 milliards de paramètres et a été pré-entraîné sur 300 milliards de *tokens*.

Contrairement aux modèles type BERT où l'affinage du modèle ne s'effectue que sur une seule tâche cible, les modèles GPT permettent l'affinage en multi-tâches en combi-

nant la tâche de modélisation (*Causal Language Modeling*) avec la tâche cible. À partir de GPT-2, le problème d'apprentissage multi-tâches est traité de manière non supervisée, où chaque tâche est vue comme un exemple de pré-entraînement. L'objectif est donc de rassembler des corpus variés afin que le modèle soit capable de généraliser. Afin d'avoir un corpus d'apprentissage de qualité, les auteurs de GPT-2 ont opté pour un corpus de pages web, WebText (Radford et al., 2019), collectées à partir de liens sortants du réseau social Reddit. Ayant été validé par plusieurs humains, ces pages web peuvent être considérées de qualité pour le corpus. Après quelques modifications sur l'architecture de GPT pour optimiser davantage le modèle, GPT-2 est entraîné sur le corpus WebText. Le modèle est ensuite appliqué à d'autres tâches grâce à un conditionnement lors de l'inférence, permettant ainsi d'évaluer le modèle en *zero-shot* (c'est-à-dire évaluer le modèle sans faire d'affinage au préalable). Ce conditionnement (aussi appelée « amorce » ou « *prompt* ») prend la forme d'un ajout d'informations au début de la séquence pour indiquer au modèle la tâche à réaliser. Par exemple, pour la tâche de traduction de l'anglais vers le français, une phrase conditionnée pourrait être exprimée de la manière suivante : « `Translate from English to French </s> The patient has a history of myocardial infarction.` ».

Ce mécanisme de conditionnement a été approfondi avec le modèle GPT-3 (Brown et al., 2020). Celui-ci reprend l'architecture de GPT-2 et a été entraîné sur un corpus composé de plusieurs sources : (i) une version de Common Crawl filtrée (410 milliards de *tokens*), (ii) WebText2, une version étendue de WebText (19 milliards de *tokens*), (iii) deux corpus de livres disponibles sur Internet (Books1 et Books2, 12 milliards et 55 milliards de *tokens* respectivement) et (iv) Wikipedia en anglais. GPT-3 a été conçu dans le but explicite d'utiliser le conditionnement, aussi appelé *prompting*, comme moyen de résoudre des tâches. Le modèle est ensuite évalué selon plusieurs stratégies :

- Affinage (ou *Fine-tuning*) : consiste à reprendre les poids du modèle pré-entraîné et à continuer l'entraînement de manière supervisée sur le jeu d'apprentissage de la tâche cible, le modèle est ensuite évalué sur le jeu de test. GPT-3 n'a pas été évalué en affinage afin de pouvoir focaliser l'évaluation du modèle sur les performances agnostiques à la tâche.

- *Few-Shot* : consiste à donner quelques exemples de la tâche à réaliser au modèle, mais sans continuer l'entraînement des poids du modèle sur ces exemples, comme illustré dans la figure 2.10.

- *One-Shot* : identique au *Few-Shot*, à l'exception qu'un seul exemple est donné au

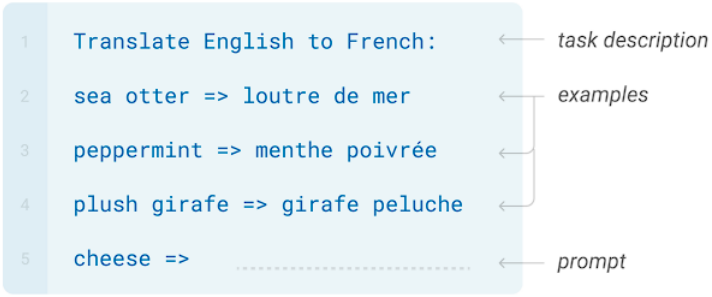


FIGURE 2.10 : Illustration du *Few-Shot*. Le modèle prédit la réponse à partir d'une description en langage naturel de la tâche et de quelques exemples. Aucune mise à jour des poids du modèle n'est effectuée. Figure extraite de l'article de Brown et al. (2020).

modèle en plus de la description de la tâche en langage naturel. Un exemple de *One-Shot* est donné dans la figure 2.11.


FIGURE 2.11 : Illustration du *One-Shot*. Le modèle prédit la réponse à partir d'une description en langage naturel de la tâche et d'un seul exemple. Aucune mise à jour des poids du modèle n'est effectuée. Figure extraite de l'article de Brown et al. (2020).

- *Zero-Shot* : seule une instruction décrivant la tâche est donnée au modèle, comme illustré dans la figure 2.12.


FIGURE 2.12 : Illustration du *zero-shot*. Le modèle prédit la réponse à partir d'une description en langage naturel de la tâche et d'un seul exemple. Aucune mise à jour des poids du modèle n'est effectuée. Figure extraite de l'article de Brown et al. (2020).

Les modèles GPT et les méthodes de conditionnement ont montré que la tâche de modélisation du langage (*Causal Language Modeling*) pouvait être directement appliquée aux tâches cible et ont ainsi inspiré de nouvelles méthodes. La grande taille des modèles

et leur capacité à être conditionnés par un *prompt* ont conduit à l'affinage à base de *prompt* (ou *prompt tuning*) où de nouvelles représentations spéciales pour les *prompts* sont entraînés pour adapter le modèle à la tâche sans modifier les paramètres du modèle, donnant comme résultat les modèles à instructions comme ChatGPT [1].

L'apparition soudaine de nombreux modèles de langue auto-régressif de plus en plus grands a amené la communauté à réfléchir à la façon de choisir le nombre de paramètres et la quantité de données optimales pour un modèle. Pour cela, Hoffmann et al. (2022) ont introduit la loi d'échelle Chinchilla (*Chinchilla scaling law*) énonçant que pour entraîner un modèle de langue auto-régressif à partir d'un certain budget en FLOPs[2], le nombre de paramètres du modèle ($N$) et le nombre de *tokens* utilisés pour entraîner le modèle ($D$) devraient augmenter à peu près dans des proportions égales afin d'atteindre une optimisation de calcul. Cette conclusion diffère de la loi d'échelle précédente pour les modèles de langage neuronaux (Kaplan et al., 2020) qui stipule que $N$ devrait augmenter plus rapidement que $D$.

Suite à GPT-3 et la loi d'échelle introduite avec le modèle Chinchilla, de nombreux grands modèles de langue ont vu le jour, telles que BLOOM (Scao et al., 2022), PaLM (Chowdhery et al., 2022) puis PaLM 2 (Anil et al., 2023), LLaMA (Touvron et al., 2023a) puis LLaMA-2 (Touvron et al., 2023b) ou encore GPT-4 (OpenAI, 2023). Chacun de ces modèles auto-régressifs se voit systématiquement associer une version affinée du modèle pour des usages en *zero-shot* avec des instructions : BloomZ (Muennighoff et al., 2023), Flan-PaLM (Chung et al., 2022), LLaMA-2-Chat (Touvron et al., 2023b) et ChatGPT.

### 2.4.5 Modèle Encodeur-Décodeur : T5 et BART

Les modèles Encodeur-Décodeur reprennent l'architecture initiale du *Transformer*. L'encodeur prend en entrée le texte source et le transforme en une représentation vectorielle, aussi appelé « contexte ». L'encodeur capte les informations sémantiques et structurelles du texte source. Le décodeur prend ensuite la représentation vectorielle générée par l'encodeur et la transforme en texte cible. Le décodeur génère le texte cible en prenant en compte le contexte encodé ainsi que les *tokens* qu'il a déjà généré. Ce type de modèle est généralement utilisé pour la traduction automatique, la génération de texte, le résumé de texte, etc.

1. `https://chat.openai.com/`
2. en anglais *floating-point operations per second*, correspond au nombre d'opérations en virgule flottante par seconde

T5 (Raffel et al., 2020) est un modèle Transformer encodeur-décodeur pré-entraîné sur des tâches non supervisées et supervisées, dans lequel chaque tâche supervisée est convertie en un format texte-à-texte. La conversion des tâches supervisées reprend le principe de *prompting* introduit dans la section 2.4.4 avec les modèles GPT-2 et GPT-3. Un préfixe est défini pour chaque tâche puis ajouté au début de chaque texte en entrée. Par exemple, pour la tâche de résumé, le préfixe est « `summarize:` » et pour la tâche d'analyse de sentiment, le préfixe est « `sst2 sentence:` ». Plusieurs corpus incluant plusieurs tâches ont donc été utilisés. Pour les tâches non supervisées, 750 GB ont été extraits de l'archive Common Crawl pour construire le corpus C4 (*Colossal Clean Crawled Corpus*) de pré-entraînement. Pour les tâches supervisées, toutes les tâches des benchmarks GLUE (Wang et al., 2018b) et SuperGLUE (Wang et al., 2019a) ont été utilisées. À partir du corpus C4 et des tâches supervisées converties au format texte-à-texte, le modèle est pré-entraîné de manière auto-supervisé avec la tâche de remplacement de segment (*span replacement*), inspirée du *Masked Language Modeling* de BERT. Le remplacement de segments consiste à échantillonner aléatoirement puis supprimer 15 % des *tokens* dans la séquence d'entrée. Tous les *tokens* consécutifs supprimés sont remplacés par un *token* sentinel unique, contrairement au *Masked Language Modeling* où chaque *token* est individuellement remplacé par le *token* `<MASK>`. La figure 2.13 donne un exemple de (*span replacement*). Lors de l'apprentissage, l'entrée de l'encodeur est la phrase avec les segments remplacés par des *tokens* sentinel (*Inputs*), l'entrée du décodeur est la phrase originale (*Original text*) et la cible est ensuite constituée des *tokens* supprimés, délimités par leurs *tokens* sentinelles associés (*Targets*).

Deux versions de T5 existent : une version avec uniquement le pré-entraînement sur le corpus C4 (T5v1.1) et une version avec le pré-entraînement sur le corpus C4 et les tâches des benchmarks GLUE et SuperGLUE. D'autres versions de T5 ont été ensuite introduites, dont une version multilingue incluant 101 langues (mT5 (Xue et al., 2021)), une version entraînée avec différentes tâches de pré-entraînement (UL2 (Tay et al., 2022)), sans oublier les versions *prompt* (ou instructions) de chacun de ces modèles (Flan-T5 (Chung et al., 2022), Flan-UL2 (Tay et al., 2022)). Tous les modèles T5 et leurs dérivées sont disponibles en plusieurs tailles, allant de quelques millions de paramètres à plusieurs milliards.

D'autres modèles reprennent également l'architecture originale du *Transformer*, comme BART (Lewis et al., 2020), un modèle séquence-à-séquence (*Sequence-to-sequence*) dont l'architecture est composée d'un encodeur bidirectionnel (similaire à BERT) et un décodeur de gauche à droite (comme GPT). Le modèle est pré-entraîné avec une tâche objectif

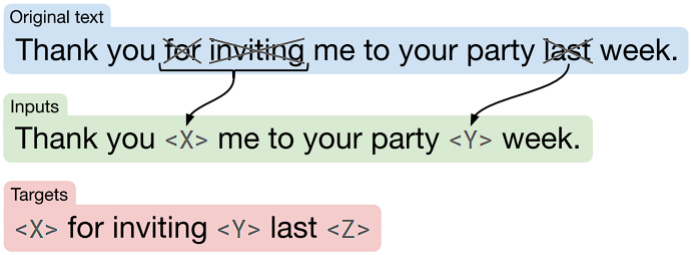


FIGURE 2.13 : Illustration de la tâche objectif de remplacement de segments (*Span Replacement*). Les mots « *for* », « *inviting* » et « *last* » sont choisis aléatoirement pour être corrompus. Chaque séquence de *tokens* corrompus est remplacée par un *token* sentinel unique (il y a autant de *tokens* sentinel que de séquence de *tokens* corrompus dans le texte). Les mots « *for* » et « *inviting* » sont consécutifs et donc remplacés par un seul *token* sentinel. La séquence de sortie est composée des séquences supprimées, délimités par le *token* sentinel auquel chaque séquence correspond. Un *token* sentinel final `<Z>` est ajouté pour marquer la fin de la séquence. Figure extraite de l'article de Raffel et al. (2020).

combinant le remplissage de texte (*text infilling*) et la permutation aléatoire de l'ordre des phrases (*sentence permutation*). La tâche de remplissage de texte consiste à un échantillon d'étendues de texte, où la longueur de chaque séquence de texte est extraite d'une distribution de Poisson ($\lambda = 3$). Chaque séquence est ensuite masquée avec un seul *token* `<MASK>`. Concernant les caractéristiques de l'architecture de BART, le modèle est composé de 12 couches dans l'encodeur et le décodeur, une taille de sortie de 1 024. BART utilise le même algorithme BPE que GPT-2 et a été pré-entraîné sur les mêmes données que RoBERTa.

### 2.4.6 Adaptation de modèles pré-entraînés à des domaines spécialisés

Les architectures et modèles de langue présentés dans la section précédente sont systématiquement pré-entraînés sur des données en anglais issues du domaine dit « général », comme des articles de presse, des livres, des articles scientifiques ou encore des encyclopédies. L'anglais est largement utilisé comme langue de communication internationale, ce qui en fait la source de données la plus disponible sur le web, facilitant ainsi la construction de modèles sur cette langue. Grâce aux sources de données volumineuses, ces modèles généralistes ont l'avantage de pouvoir modéliser efficacement la langue telle qu'elle est utilisée couramment. Cela garantit une robustesse de ces modèles dans la majorité des

applications ou cas d'usages du TAL. Pour les langues autres que l'anglais, la disponibilité de tels modèles monolingues est moindre, car limité par la quantité de ressources de calcul nécessaire pour le pré-entraînement (plusieurs centaines de GPUs) et surtout par la quantité de données disponibles dans la langue.

Au fil du temps, les architectures des modèles existants ont été réemployées pour pré-entraîner des modèles de langues monolingues pour d'autres langues. Par exemple, pour le français, CamemBERT (Martin et al., 2020) et FlauBERT (Le et al., 2020) ont été les deux premiers modèles proposés. Le modèle CamemBERT reprend l'architecture de RoBERTa mais utilise l'algorithme SentencePiece pour construire son vocabulaire d'une taille de 32 000 tokens. Le modèle est ensuite pré-entraîné sur la partie française du corpus OSCAR (Ortiz Suárez et al., 2019) d'environ 138 GB avec la tâche de *Whole-Word Masking* dynamique. Le modèle FlauBERT reprend l'architecture de BERT mais utilise l'algorithme BPE avec une taille de vocabulaire de 50 000 tokens. Le modèle est ensuite pré-entraîné sur la tâche de *Masking Language Modeling* dynamique sur un corpus constitué de trois sources : WMT19 (Li et al., 2019), la collection OPUS (Tiedemann, 2012) et des jeux de données disponibles dans le projet Wikimedia, pour un total de 71 GB de données. Plus récemment, le modèle CamemBERTa (Antoun et al., 2023a) a été introduit, celui-ci reprenant l'architecture de DeBERTaV3 (He et al., 2021a) et pré-entraîné sur le corpus CCNet (Wenzek et al., 2020).

Bien que les modèles généraux soient polyvalents et robustes, leur usage dans des domaines plus spécialisés comme le domaine médical est plus limité. En effet, les domaines spécialisés ont leurs propres spécificités lexicales et sémantiques s'éloignant du domaine général. Les termes très présents dans les domaines spécialisés sont moins bien représentés par les modèles généraux. En ce sens, plusieurs travaux ont montré qu'il est nécessaire d'adapter les modèles généraux au domaine spécialisé cible afin d'obtenir les meilleures performances (El Boukkouri et al., 2022; Lee et al., 2019; Chalkidis et al., 2020; Li et al., 2023).

L'adaptation à un domaine peut être réalisée à l'aide de plusieurs méthodes (Ramponi and Plank, 2020), celles-ci étant réparties en trois catégories : les méthodes centrées sur les modèles (*model-centric*) qui favorisent les approches modifiant les caractéristiques du modèle, comme sa fonction objectif, son architecture ou encore ses poids ; les méthodes centrées sur les données (*data-centric*) qui se concentrent sur la sélection de données représentatives du domaine de spécialité en utilisant un modèle source pour générer des données étiquetées (i.e. *pseudo-labeling*), des grands corpus non étiquetés ou des tâches

auxiliaires ayant des données étiquetées (i.e. pré-entraînement) ; et les méthodes hybrides combinant les méthodes centrées sur les modèles et les données. Pour la suite, nous nous focalisons sur les méthodes centrées sur les données, et plus particulièrement l'adaptation par pré-entraînement, puisque ce sont les méthodes que nous avons explorées dans ce manuscrit.

L'adaptation par pré-entraînement repose sur un corpus issu du domaine sur lequel nous souhaitons spécialiser le modèle. Le pré-entraînement de modèles de langue suit généralement deux stratégies : le pré-entraînement de zéro (*from scratch*) et le pré-entraînement poursuivi (*continual pre-training*).

**Pré-entraînement de zéro (*from scratch*)** Le pré-entraînement de zéro consiste à entraîner un modèle uniquement sur des données du domaine spécialisé cible. Le vocabulaire du modèle est appris sur le corpus spécialisé, puis le modèle est pré-entraîné sur ce même corpus. Si le pré-entraînement de zéro était initialement pensé pour entraîner des modèles à apprendre et à exploiter des caractéristiques permettant un usage général du modèle, il est aussi utilisable pour entraîner les modèles spécialisés lorsque la quantité de données du domaine cible est suffisamment élevée (Beltagy et al., 2019; Gu et al., 2021).

**Pré-entraînement poursuivi (*continual pre-training*)** Le pré-entraînement poursuivi consiste à reprendre un modèle existant et continuer son pré-entraînement sur des données du domaine spécialisé cible généralement de plus petites tailles. Le vocabulaire du modèle de langue général est utilisé pour segmenter le corpus spécialisé. Cette approche consiste en deux pré-entraînements consécutifs : un pré-entraînement initial qui apprend et transfère des caractéristiques générales, puis, un second pré-entraînement spécifique au domaine qui exploite les caractéristiques générales apprises pour les transférer vers le domaine cible. Pré-entraîner un modèle plusieurs fois sur des corpus issus de domaines différents s'inscrit dans les méthodes de pré-entraînement adaptatif à phases multiples (*multi-phase adaptative pre-training*) (Ramponi and Plank, 2020). Parmi les méthodes de pré-entraînement adaptatif à phases multiples, deux méthodes sont souvent comparées : le pré-entraînement adaptatif au domaine (*domain-adaptative pre-training* ou DAPT) (Dodge et al., 2019) à partir d'un corpus large propre au domaine et le pré-entraînement spécifique à la tâche (*task-specific pre-training* ou TAPT) (Han and Eisenstein, 2019) qui utilise des données non étiquetées proches de la distribution de la tâche, par exemple, continuer le pré-entraînement d'un modèle sur le sous-ensemble d'apprentis-

sage d'une tâche pour ensuite affiner le modèle sur la tâche cible. Ces études ont montré l'importance du pré-entraînement sur des données spécialisées, quelle que soit la quantité de données, et que l'utilisation conjointe des deux méthodes (DAPT et TAPT) améliore encore les performances. L'utilisation consécutive du DAPT et du TAPT donne un triple pré-entraînement : un pré-entraînement du modèle de langue sur le domaine général, puis un pré-entraînement spécifique au domaine à l'aide d'un grand corpus propre au domaine, et enfin un pré-entraînement spécifique à la tâche en utilisant le corpus de la tâche cible.

Les pré-entraînements de zéro et poursuivi peuvent être employés pour n'importe quels types de modèles (encodeur, decodeur, encoder-decodeur). À partir de ces deux grandes stratégies d'adaptation de modèle de langues, beaucoup de modèles ont été adaptés pour les domaines biomédical et clinique. Nous entendons par domaine biomédical les modèles entraînés sur des données issues de sources libres, comme l'UMLS ou les bases de données bibliographiques telles que PubMed ou MEDLINE, et par domaine clinique les modèles entraînés sur des données cliniques issues du soin. À l'instar du domaine général, les premiers modèles pour les domaines biomédical et clinique ont été pré-entraînés sur des corpus en anglais.

Pour le domaine biomédical, la majorité des modèles existants sont des variantes de BERT. BioBERT (Lee et al., 2019) et PubMedBERT (Gu et al., 2021) sont pré-entraînés sur PudMed, le premier en pré-entraînement poursuivi sur des résumés d'articles et des articles intégraux de PubMed Central (PMC) et le second en pré-entraînement de zéro uniquement sur des résumés d'articles. D'autres modèles pré-entraînés sur différentes sources de données ont aussi été introduits. C'est le cas de SciBERT (Beltagy et al., 2019), pré-entraîné sur les articles intégraux de biomédical et d'informatique issus de Semantic Scholar, et BlueBERT (Peng et al., 2019a), pré-entraîné sur une combinaison de données biomédicales (résumés d'articles PubMed) et cliniques (base de données MIMIC-III). Concernant les modèles génératifs, SciFive (Phan et al., 2021) est une version adaptée du modèle T5, pré-entraîné sur des résumés d'articles PubMed et des articles intégraux de PMC. Une version de BART a aussi été proposée, BioBART (Yuan et al., 2022), pré-entraîné sur des résumés d'articles PubMed. Plus récemment, les larges modèles de langue se voient aussi adaptés au domaine biomédical. PMC-LLaMA (Wu et al., 2023) reprend le modèle LLaMA dont le pré-entraînement a été poursuivi sur 4,8 millions d'articles universitaires biomédicaux. Le modèle LLaMA a aussi été adapté à l'aide d'instructions,

formant ainsi le modèle medAlpaca (Han et al., 2023). De la même façon, les modèles PaLM et PaLM 2 ont été adaptés avec des instructions pour le domaine médical, donnant les modèles Med-PaLM (Singhal et al., 2023a) et Med-PaLM 2 (Singhal et al., 2023b) respectivement.

Pour les langues autres que l'anglais, des modèles de langue biomédicaux ont aussi été proposés, en moindre mesure. Par exemple, pour l'allemand avec le modèle medBERT.de (Bressem et al., 2023), pour le portugais avec BioBERTpt (Schneider et al., 2020), pour l'espagnol avec biomedical-es (Carrino et al., 2021) et pour le turc avec Bioberturk (Turkmen et al., 2022). Enfin, pour le français, plusieurs modèles ont été introduits récemment, dont AliBERT (Berhe et al., 2023), CamemBERT-Bio (Touchent et al., 2023) ou encore les modèles DrBERT (Labrak et al., 2023a) et DrLongformer qui sont l'une des contributions de cette thèse et que nous présentons en détails dans le chapitre 4. Le tableau 2.1 propose une synthèse des principaux modèles de langue existants et de leurs caractéristiques pour le domaine biomédical.

| **Modèle** | **Langue** | **Architecture** | **Tokenizer** | **Pré-entraînement** | **Corpus Pré-entraînement** | **Taille** |
|---|---|---|---|---|---|---|
| BioBERT | Anglais | BERT | WordPiece 30k | Continué (BERT) | PubMed | 4,5B |
| PubMedBERT | Anglais | BERT | WordPiece 30k | De zéro | PubMed | 3,1B |
| SciBERT | Anglais | BERT | WordPiece 30k | De zéro | PMC + Semantic Scholar | 3,2B |
| BlueBERT | Anglais | BERT | WordPiece 30k | Continué (BERT) | PubMed + MIMIC-III | 4,5B |
| SciFive | Anglais | T5 | SentencePiece 32k | Conitnué (T5) | PubMed + PMC | - |
| BioBART | Anglais | BART | BPE 50k | Continué (BART) | PubMed | 41 GB |
| PMC-LLaMA | Anglais | LLaMA | SentencePiece BPE 32k | Continué (LLaMA) | PMC | 75B |
| medAlpaca | Anglais | LLaMA | SentencePiece BPE 32k | Continué (LLaMA) | Medical Meadow (Q&A) | - |
| Med-PaLM | Anglais | PaLM | SentencePiece 256k | Continué (PaLM) | Corpus Q&A médicaux | - |
| Med-PaLM 2 | Anglais | PaLM 2 | SentencePiece 256k | Continué (PaLM 2) | Corpus Q&A médicaux | - |
| medBERT.de | Allemand | BERT | WordPiece 30k | De zéro | Web crawl médical et EHR | 1,1B |
| BioBERTpt | Portugais | BERT | WordPiece 30k | Continué (BERT) | EHR + Scielo + PubMed | 44M |
| biomedical-es | Espagnol | RoBERTa | BPE 52k | De zéro | Web crawl médical | 968M |
| Bioberturk | Turc | BERT | WordPiece 32k | Continué (BERTurk) | Articles + Thèses Radiologie | 75M |
| AliBERT | Français | RoBERTa | Unigram 40k | De zéro | Web crawl médical | 7 GB |
| CamemBERT-Bio | Français | RoBERTa | SentencePiece 32k | Continué (CamemBERT) | biomed-fr | 400M |
| DrBERT | Français | RoBERTa | SentencePiece 32k | De zéro | NACHOS | 1,1B |
| DrBERT-4096 | Français | Longformer | SentencePiece 32k | Continué (DrBERT) | NACHOS | 1,1B |
| DrLongformer-FS | Français | Longformer | BPE 50k | De zéro | NACHOS | 1,1B |
| DrLongformer-CP | Français | Longformer | BPE 50k | Continué (Clinical-Longformer) | NACHOS | 1,1B |

TABLEAU 2.1 : Modèles de langues pour le domaine biomédical. La *Taille* correspond à la taille du corpus de pré-entraînement, parfois indiqué en nombre de mots (**M** pour millions et **B** pour milliards) ou en gygabytes (**GB**).

Pour le domaine clinique, la disponibilité des modèles de langues est limitée, car ceux-ci sont généralement gardés privés pour des raisons de confidentialité liés aux données de santé. La base de données MIMIC est la principale source de données cliniques utilisée pour pré-entraîner les modèles anglais. ClinicalBERT (Huang et al., 2019) et deux modèles ClinicalT5 (Lu et al., 2022; Lehman et al., 2023) ont utilisé MIMIC comme corpus de pré-entraînement. Pour les longs documents, les modèles Longformer et BigBird ont

également été adaptés avec MIMIC fournissant les modèles Clinical-Longformer (Li et al., 2023) et Clinical-BigBird (Li et al., 2023) respectivement. D'autres modèles intègrent des données cliniques (MIMIC ou autre) mais en plus petite quantité et souvent diluées dans des données biomédicales, tels que les modèles BlueBERT (Peng et al., 2019b), medBERT.de (Bressem et al., 2023) pour l'allemand ou encore BioBERTpt (Schneider et al., 2020) pour le portugais.

D'autres modèles pré-entraînés uniquement sur des données cliniques sont présentés dans la littérature mais ne sont pas rendus publics. Par exemple, l'AP-HP a pré-entraîné plusieurs modèles sur 21 millions de comptes rendus cliniques issus de leur entrepôt de données de santé (Dura et al., 2022). Un modèle en suédois, SweDeClin-BERT (Vakili et al., 2022), a été pré-entraîné sur des documents cliniques désidentifiés automatiquement. Enfin, le modèle ChuBERT (Labrak et al., 2023a) a été pré-entraîné sur des comptes rendus cliniques issus de l'entrepôt de données biomédicales du CHU de Nantes. Nous présentons ce modèle en détails dans le chapitre 4. Le tableau 2.2 propose une synthèse des principaux modèles de langue existants et de leurs caractéristiques pour le domaine clinique.

| **Modèle** | **Langue** | **Architecture** | **Tokenizer** | **Pré-entraînement** | **Corpus Pré-entraînement** | **Taille** |
|---|---|---|---|---|---|---|
| ClinicalBERT | Anglais | BERT | WordPiece 30k | Continué (BERT) | MIMIC-III | 500M |
| ClinicalT5 (1) | Anglais | T5 | SentencePiece 32k | Continué (SciFive) | MIMIC-III | 2M doc |
| ClinicalT5 (2) | Anglais | T5 | SentencePiece 32k | De zéro | MIMIC-III + MIMIC-IV | 40B |
| Clinical-Longformer | Anglais | Longformer | BPE 50k | Continué (Longformer) | MIMIC-III | 2M doc |
| Clinical-BigBird | Anglais | BigBird | SentencePiece 32k | Continué (BigBird) | MIMIC-III | 2M doc |
| AP-HP EDS-fs | Français | RoBERTa | SentencePiece 32k | Continué (CamemBERT) | Documents AP-HP | 21M doc |
| AP-HP EDS-ft | Français | RoBERTa | SentencePiece 32k | De zéro | Documents AP-HP | 21M doc |
| SweDeClin-BERT | Suédois | BERT | SentencePiece 50k | Continué (KB-BERT) | Health Bank | 50M phrases |
| ChuBERT | Français | RoBERTa | SentencePiece 32k | De zéro | Documents Chu Nantes | 655M |

Tableau 2.2 : Modèles de langues pour le domaine clinique. La *Taille* correspond à la taille du corpus de pré-entraînement, parfois indiqué en nombre de mots (**M** pour millions et **B** pour milliards), en gygabytes (**GB**) ou en nombre de documents (**doc**.

## 2.5 Évaluation extrinsèque en domaine biomédical et clinique

L'évaluation des modèles de langue est une étape essentielle pour vérifier leur qualité et leurs performances. Cette validation consiste généralement à évaluer leurs performances sur des tâches cibles spécifiques. La disponibilité de corpus annotés joue donc un rôle crucial puisqu'ils fournissent une vérité terrain qui sera ensuite confrontée aux prédictions des modèles pour mesurer précisément leurs performances. L'évaluation des modèles nécessite des corpus sur des tâches et des sources de données variées, à l'instar de ce qui est disponible pour le domaine général comme GLUE (Wang et al., 2018b).

Les corpus présents dans la littérature sont composés de différentes sources de données. D'un côté, les corpus biomédicaux s'appuient sur des sources de données ouvertes, comme l'UMLS ou les bases de données bibliographiques (PubMed, MEDLINE) avec les résumés d'articles scientifiques et les cas cliniques. De l'autre côté, les corpus cliniques s'appuient sur des bases de données cliniques issues du soin. Pour l'anglais, MIMIC est la source principalement utilisée pour la création de corpus cliniques grâce aux documents cliniques désidentifiés accessibles dans la base de données. En dehors des corpus provenant de MIMIC, les corpus cliniques accessibles pour la recherche sont plutôt rares, voire inexistants pour la majorité des langues autres que l'anglais. Nous proposons ci-dessous un tour d'horizon des tâches de TAL que l'on peut retrouver dans les domaines biomédical et clinique, illustrées par quelques corpus pour chaque tâche.

### 2.5.1 Reconnaissance d'entités nommées

L'objectif principal de la tâche de reconnaissance d'entités nommées est d'extraire des informations pertinentes au sein d'un texte en identifiant des entités spécifiques. Les entités recherchées peuvent varier selon la nature des documents et appartenir à différentes catégories d'entités. Parmi ces catégories d'entités, nous pouvons retrouver les maladies ou pathologies - **NCBI Disease** (Doğan et al., 2014), **BC5CDR** (Li et al., 2016), **n2c2 2010** (Uzuner et al., 2011) ; les médicaments et les composants chimiques - **BC5CDR** (Li et al., 2016), **BC4CHEMD** (Krallinger et al., 2015) ; les gènes et les protéines - **BC2GM** (Smith et al., 2008), **JNLPBA** (Collier and Kim, 2004) ; les noms d'espèce - **LINNAEUS** (Gerner et al., 2010), **Species-800** (Pafilis et al., 2013) ; les entités temporelles — **n2c2 2012** (Sun et al., 2013) ; ou encore les entités pour la dési-

dentification — **n2c2 2006** (Uzuner et al., 2007), **i2b2 2014 Track 1** (Stubbs et al., 2015). Pour les corpus français, on retrouve les entités cliniques — **E3C** (Magnini et al., 2020), **DEFT-2019** (Cardon et al., 2020), **DEFT-2021** (Grouin et al., 2021) ; de négation et d'incertitude — **CAS corpus** (Grabar et al., 2018), **ESSAI** (Dalloux et al., 2021) ; de groupes sémantiques UMLS — **QUAERO** (Névéol et al., 2014), **Mantra-GSC** (Kors et al., 2015) ; de prescriptions de médicaments — **PxCorpus** (Kocabiyikoglu et al., 2022) et l'étiquetage morpho-syntaxique - **CRTT-MED** (Maniez, 2011). Le tableau 2.3 synthétise les principales informations sur ces corpus.

| **Corpus** | **Source** | **# Instances** | **# Mots moyen** | **Langue** | **Accès** |
|---|---|---|---|---|---|
| BC2GM | MEDLINE | 20 000 | 28,49 | EN | Public |
| BC4CHEMD | PubMed | 87 688 | 29,01 | EN | Public |
| BC5CDR | PubMed | 1 500 | 175,81 | EN | Public |
| i2b2 2014 Track 1 | CR cliniques | 1 304 | 914,50 | EN | Accord utilisation |
| JNLPBA | MEDLINE | 44 808 | 26,50 | EN | Public |
| LINNAEUS | PubMed | 84 | 23,29 | EN | Public |
| n2c2 2006 | CR cliniques | 889 | 3605,62 | EN | Accord utilisation |
| n2c2 2010 | CR cliniques | 426 | 5403,48 | EN | Accord utilisation |
| n2c2 2012 | CR cliniques | 310 | 532,41 | EN | Accord utilisation |
| Species-800 | PubMed | 8 194 | 25,84 | EN | Public |
| NCBI Disease | PubMed | 792 | 179,92 | EN | Public |
| CAS corpus | Cas cliniques | 3 790 | 23,07 | FR | Accord utilisation |
| ESSAI | Cas cliniques | 7 247 | 22,55 | FR | Accord utilisation |
| QUAERO | Notices de médicaments et titres d'articles biomédicaux | 2 536 | - | FR | Public |
| E3C | Cas cliniques | - | - | FR | Public |
| Mantra-GSC | Résumés et titres d'articles biomédicaux, médicaments et brevets | 250 | 26,61 | FR | Public |
| DEFT-2019 | Cas cliniques | 717 | 26,61 | FR | Accord utilisation |
| DEFT-2021 | Cas cliniques | 275 | - | FR | Accord utilisation |
| PxCorpus | Transcriptions de prescriptions de médicaments | 1 981 | 11,33 | FR | Accord utilisation |
| CRTT-MED | Articles Science direct | 663 | - | FR | Public |

TABLEAU 2.3 : Exemples de corpus biomédicaux et cliniques sur la reconnaissance d'entités nommées.

### 2.5.2 Extraction de relations

La tâche d'extraction de relations consiste à extraire les relations sémantiques d'intérêt entre les entités dans un texte. Dans les domaines biomédical et clinique, l'extraction de relations peut être utilisée pour détecter les relations entre les gènes et les maladies - **GAD** (Bravo et al., 2015a), **EU-ADR** (Van Mulligen et al., 2012) ; entre les composants chimiques et les protéines - **CHEMPROT** (Krallinger et al., 2017) ; entre les problèmes médicaux, les examens et les traitements — **i2b2 2010** (Sun et al., 2013) ; entre les évènements cliniques et les expressions temporelles — **n2c2 2012** (Sun et al., 2013) ; ou encore les relations d'équivalence entre deux entités - **GENIA Relation Corpus** (Kim

et al., 2003). Le tableau 2.4 synthétise les principales informations sur ces corpus.

| Corpus | Source | # Instances | # Mots moyen | Langue | Accès |
|---|---|---|---|---|---|
| CHEMPROT | PubMed | 2 432 | 225,42 | EN | Public |
| EU-ADR | MEDLINE | 300 | 199,90 | EN | Public |
| GAD | MEDLINE | 5 330 | 26,57 | EN | Public |
| GENIA | GENIA | 1 210 | 220,85 | EN | Public |
| n2c2 2010 | CR cliniques | 426 | 5 403,48 | EN | Accord utilisation |
| n2c2 2012 | CR cliniques | 300 | 532,41 | EN | Accord utilisation |

Tableau 2.4 : Exemples de corpus biomédicaux ou cliniques sur l'extraction de relations.

### 2.5.3 Classification de texte

La tâche de classification de texte consiste à attribuer une étiquette à un texte donné. Le texte peut être de différentes nature et longueur, tel qu'un document, un paragraphe ou une phrase. Plusieurs types de classification sont possibles selon le nombre d'étiquettes à attribuer au texte. Le cas le plus simple, la classification binaire, où une étiquette est attribuée à un texte parmi un choix de deux étiquettes. Dans le domaine biomédical et clinique, la classification binaire peut porter sur des dialogues entre patients et médecins pour identifier le locuteur - **Meddialog** (Chen et al., 2020b) ; des échantillons provenant d'expériences de microréseaux, de transcriptomique et de cellules uniques à classer en tant que contrôle ou perturbation - **GEOKhoj v1** (Elucidata, 2020) ; la présence d'associations entre gènes et maladies dans des études de génétique - **GAD** (Bravo et al., 2015b) ; ou encore prédire la présence d'une maladie, par exemple la possibilité d'une lésion rénale aiguë chez un patient - **MIMIC-AKI** (Li et al., 2018; Sun et al., 2019). Ensuite, la classification multi-classes où une étiquette peut être attribuée à un texte parmi un choix de plusieurs étiquettes. Parmi les corpus existants pour cette tâche, on peut retrouver le corpus de l'édition 2006 du challenge n2c2 sur la classification du statut tabagique dans des textes cliniques (Uzuner et al., 2008a) ou le corpus **OHSUMED** (Hersh et al., 1994) sur la classification de résumés d'articles scientifiques avec les catégories du MeSH. Enfin, la classification multi-étiquettes où plusieurs étiquettes peuvent être attribuées à un texte parmi un choix de plusieurs étiquettes. Les corpus de classification multi-étiquettes sont les plus complexes car ils rassemblent souvent un nombre d'étiquettes très élevé comme le codage CIM-10 avec le corpus **CodiEsp** (Miranda-Escalada et al., 2020), les termes MeSH avec le corpus **OpenI** (Demner-Fushman et al., 2012) et les caractéristiques du cancer avec le corpus **Hallmarks of Cancer (HoC)** (Baker et al., 2016). Pour les corpus français, la

classification de textes porte sur la négation et l'incertitude - **CAS corpus** (Grabar et al., 2018), **ESSAI** (Dalloux et al., 2021) ; les spécialités médicales - **MorFITT** (Labrak et al., 2023b) ; la similarité entre phrases - **DEFT-2020** (Cardon et al., 2020) ; les chapitres du MeSH - **DEFT-2021** (Grouin et al., 2021) et l'intention - **PxCorpus** (Kocabiyikoglu et al., 2022). Le tableau 2.5 synthétise les principales informations sur ces corpus.

| **Corpus** | **Tâche** | **Source** | **# Instances** | **# Mots moyen** | **Langue** | **Accès** |
|---|---|---|---|---|---|---|
| CodiEsp | Multi-étiquettes | Cas cliniques | 1 000 | 346,59 | ES | Public |
| GAD | Binaire | MEDLINE | 5 330 | 26,57 | EN | Public |
| GEOKhoj v1 | Binaire | Gene Expression Omnibus | 30 000 | 23,20 | EN | Public |
| Hallmarks of Cancer | Multi-étiquettes | PubMed | 17 464 | 25,84 | EN | Public |
| Meddialog | Binaire | Web : healthcaremagic et icliniq | 1 229 | 52,62 | EN,ZH | Public |
| MIMIC-AKI | Binaire | MIMIC | 16 536 | 1463,10 | EN | Accord utilisation |
| n2c2 2006 | Multi-classes | CR cliniques | 502 | 688,5 | EN | Accord utilisation |
| OHSUMED | Multi-classes | MEDLINE | 348 564 | 159,48 | EN | Public |
| OpenI | Multi-étiquettes | IndianaU | 3 684 | 69,7 | EN | Public |
| CAS corpus | Multi-classes | Cas cliniques | 3 790 | 23,07 | FR | Accord utilisation |
| ESSAI | Multi-classes | Cas cliniques | 7 247 | 22,55 | FR | Accord utilisation |
| MorFITT | Multi-étiquettes | Résumés d'articles biomédicaux | 3 624 | 226,33 | FR | Public |
| DEFT-2020 | Multi-classes | Cas cliniques, encyclopédies et médicaments | 1 102 | 90,19 | FR | Accord utilisation |
| DEFT-2021 | Multi-étiquettes | Cas cliniques | 275 | 332,55 | FR | Accord utilisation |
| DEFT-2021 | Multi-classe | Transcriptions de prescriptions de médicaments | 1 981 | 11,33 | FR | Accord utilisation |

Tableau 2.5 : Exemples de corpus biomédicaux ou cliniques sur la classification de texte.

### 2.5.4 Questions-réponses

La tâche de questions-réponses vise à développer des systèmes permettant de répondre automatiquement à des questions. Les questions-réponses peuvent être classés en deux catégories : les questions à réponses ouvertes et les questions à choix multiples. Les questions à réponse ouvertes visent à répondre à des questions plus complexes en générant du texte en langage naturel. Les questions à choix multiples se concentrent des réponses précises à des questions factuelles, soit avec des réponses binaires (oui/non ou vrai/faux), soit avec des réponses multiples, où une à plusieurs réponses sont à sélectionner parmi un choix de réponses proposées. Dans le domaine biomédical, les questions à réponse ouvertes sont souvent associés à des cas cliniques ou des dossiers patients contenant la réponse à la question posée, par exemple le corpus **emrQA** (Pampari et al., 2018). Autrement, la majorité des corpus de questions-réponses sont composés de questions à choix multiple (**BioASQ Task B** (Tsatsaronis et al., 2015), **MedQA** (Jin et al., 2021)). Le tableau 2.6 synthétise les principales informations sur ces corpus.

| Corpus | Source | # Instances | # Mots moyen | Langue | Accès |
|---|---|---|---|---|---|
| BioASQ Task B | Multiple | 12 025 | 229,69 | EN | Accord utilisation |
| emrQA | CR cliniques | 427 882 | 2 877,76 | EN | Accord utilisation |
| MedQA | National Medical Board Examination | 14 123 | 115,27 | EN | Public |

Tableau 2.6 : Exemples de corpus biomédicaux ou cliniques sur le question-réponse.

### 2.5.5 Résolution de coréférences

L'objectif de la tâche de résolution de coréférences est d'identifier et de regrouper les expressions qui font référence à la même entité dans un texte. La résolution de coréférences permet de désambiguïser les textes et d'améliorer la compréhension globale. Dans les corpus biomédicaux, la résolution de coréférences porte sur les termes médicaux — **n2c2 2011** (Uzuner et al., 2012) ; les gènes et les protéines - **BioNLP-ST 2011 REL** (Pyysalo et al., 2011) et les microorganismes - **BioNLP 2019 BB** (Bossy et al., 2019). Le tableau 2.7 synthétise les principales informations sur ces corpus.

| Corpus | Source | # Instance | # Mots moyen | Langue | Accès |
|---|---|---|---|---|---|
| BioNLP-ST 2011 REL | GENIA | 1 210 | 220,74 | EN | Public |
| BioNLP-ST 2019 BB | GENIA | 295 | 152,24 | EN | Public |
| n2c2 2011 | CR cliniques | 424 | 976,41 | EN | Accord utilisation |

Tableau 2.7 : Exemples de corpus biomédicaux ou cliniques sur la résolution de coréférences.

### 2.5.6 Similarité sémantique

La tâche de similarité sémantique consiste à mesurer le degré de similitude entre deux textes. L'objectif est de déterminer si les deux textes présentent un sens similaire ou partagent des informations semblables, même si leur formulation diffère. La similarité sémantique peut être mesurée entre des termes, tels que les termes médicaux avec les corpus **EHR-Rel** (Schulz et al., 2020), **UMNSRS** (Pakhomov et al., 2010), **MayoSRS** et **MiniMayoSRS** (Pedersen et al., 2007), les verbes ou les noms avec les corpus **Bio-SimVerb** ou **Bio-SimLex** respectivement (Chiu et al., 2018). La similarité sémantique est aussi évaluée sur des textes plus longs tels que des phrases, comme présenté dans le corpus **BIOSSES** (Soğancıoğlu et al., 2017). Pour le français, les corpus **CLISTER** (Hiebel et al., 2022) et **DEFT-2020** (Cardon et al., 2020) traitent uniquement la similarité entre phrases. Le tableau 2.8 synthétise les principales informations sur ces corpus.

| Corpus | Source | # Instance | # Mots moyen | Langue | Accès |
|---|---|---|---|---|---|
| Bio-SimLex | PubMed | 988 | 2,0 | EN | Public |
| Bio-SimVerb | PubMed | 1 000 | 2,0 | EN | Public |
| BIOSSES | Articles scientifiques | 100 | 45,98 | EN | Public |
| EHR-Rel | IQVIA Medical Research Data | 3 741 | 6,25 | EN | Public |
| MayoSrs | Mayo Clinic | 101 | 3,11 | EN | Public |
| MiniMayoSRS | Mayo Clinic | 29 | 3,17 | EN | Public |
| UMNSRS | UMLS | 587 | 2,04 | EN | Public |
| CLISTER | Cas cliniques | 1 000 | 25,21 | FR | Accord utilisation |
| DEFT-2020 | Cas cliniques, encyclopédies et médicaments | 1 010 | 45,14 | FR | Accord utilisation |

Tableau 2.8 : Exemples de corpus biomédicaux ou cliniques sur la similarité sémantique.

### 2.5.7 Reconnaissance d'inférences textuelles

La tâche de reconnaissance d'inférences textuelles (*natural language inference*) consiste à déterminer la relation sémantique entre une prémisse et une hypothèse. Plus précisément, il s'agit de déterminer si l'hypothèse peut être dérivée ou inférée de la prémisse. Pour cela, chaque paire de prémisse et hypothèse peut être classée parmi les trois relations suivantes :

- Implication : la prémisse implique l'hypothèse, c'est-à-dire que l'hypothèse est considérée vraie en fonction de la prémisse.
- Contradiction : la prémisse et l'hypothèse sont en contradiction, c'est-à-dire que l'hypothèse est considérée fausse en fonction de la prémisse.
- Neutre : la prémisse et l'hypothèse n'ont aucune relation sémantique forte, on dit dans ce cas qu'elles sont neutres.

Plusieurs corpus ont été proposés pour cette tâche, comme **MedNLI** (Romanov and Shivade, 2018), **SciTail** (Khot et al., 2018) et **Evidence Inference 2.0** (DeYoung et al., 2020). Le tableau 2.9 synthétise les principales informations sur ces corpus.

| Corpus | Source | # Instance | # Mots moyen | Langue | Accès |
|---|---|---|---|---|---|
| Evidence Inference 2.0 | Articles scientifique | 12 616 | 43,95 | EN | Public |
| MedNLI | MIMIC | 14 049 | 21,40 | EN | Accord utilisation |
| SciTail | Jeu de données SciQ | 27 026 | 28,94 | EN | Public |

Tableau 2.9 : Exemples de corpus biomédicaux ou cliniques sur la reconnaissance d'inférences textuelles.

## 2.6 Conclusion

Dans ce chapitre, nous avons défini les méthodes de représentations sémantiques des mots. Les représentations discrètes, basées principalement sur les sacs de mots, ont été les premières représentations sémantiques utilisées pour représenter les mots avec des vecteurs. Les vecteurs obtenus avec ces approches ont l'avantage d'être explicables car chaque dimension représente un concept précis. Cependant, ces vecteurs sont de grandes dimensions et creux, ce qui rend difficile leur utilisation dans des architectures neuronales où la taille des vecteurs à un impact considérable sur la complexité. Pour pallier ces inconvénients, les représentations continues et statiques ont été introduits. Ces représentations sont des vecteurs denses de faibles dimensions mais plus difficile à interpréter. Ensuite, nous avons présenté les représentations continues et contextualisées permettant d'avoir plusieurs représentations pour un même mot selon le contexte dans lequel il apparaît. Nous avons détaillé l'architecture *Transformer* et les différents modèles reposant sur celle-ci. Les méthodes permettant l'adaptation des modèles de langues à un domaine spécialisé ont également été introduites. Enfin, nous avons recensé les corpus biomédicaux et cliniques développés pour les principales tâches de TAL. Ces différentes tâches sont essentielles pour l'évaluation extrinsèque des modèles de langue.

Dans le chapitre 4 de ce manuscrit, nous présentons l'adaptation au domaine médical des modèles BERT et Longformer pour combler le manque de modèles spécialisés pour le français. Nous introduisons également, dans le chapitre 3, deux corpus cliniques constitués de comptes rendus hospitaliers du CHU de Nantes pour une évaluation extrinsèque des modèles sur des tâches cliniques utiles.

Chapitre 3

# Construction de corpus

Les approches de TAL actuelles reposent principalement sur des algorithmes d'apprentissage profond, que ce soit en apprentissage supervisé ou non supervisé. Dans le cas des approches supervisées, les ressources de références (*gold standard*) annotées manuellement sont indispensables pour leur développement et leur évaluation. Cependant, l'annotation est une tâche très coûteuse et chronophage, surtout dans les domaines de spécialité où une expertise pointue est requise. C'est un travail qui demande une grande rigueur aux annotateurs en vue de maintenir une annotation cohérente et de qualité.

Dans le domaine médical, les corpus annotés accessibles à la communauté sont rares comme nous avons pu le décrire dans le chapitre 2 sur l'état de l'art des corpus existants. Ces corpus s'appuient en grande partie sur des données déjà accessibles sur le web comme des articles scientifiques, des cas cliniques ou des données issues des réseaux sociaux. Les corpus composés de données cliniques sont plus nombreux mais ne sont jamais rendus publics, car il est difficile de justifier du caractère anonyme d'une donnée clinique, même désidentifié ou pseudonymisé. Lorsque des approches de TAL sont employées dans les établissements de santé, notamment dans le cas d'extraction d'informations cliniques précises, l'annotation d'un corpus pour la tâche visée est incontournable. Au-delà d'être nécessaire pour développer et évaluer les approches de TAL, l'annotation d'un corpus sert dans un premier temps à vérifier si les informations d'intérêt sont bien présentes dans les documents. Une fois la présence de l'information confirmée, l'annotation sert ensuite à évaluer la qualité des informations.

Dans ce chapitre, nous présentons deux des corpus constitués dans cette thèse, tous deux composés de documents cliniques issus de l'entrepôt de données du CHU de Nantes. Dans la section 3.2, nous présentons le corpus annoté dans le cadre du projet GAVROCHE. Ce corpus est composé de comptes rendus hospitaliers et annoté avec des informations cliniques permettant de phénotyper les patients hospitalisés pour insuffisance cardiaque aiguë (ICA). Nous présentons ensuite, dans la section 3.3, un corpus composé de paragraphes issus de comptes rendus hospitaliers du CHU de Nantes où nous nous intéressons

aux déterminants sociaux de santé à travers une tâche d'extraction d'entités nommées et de relations. Enfin, nous concluons ce chapitre avec une synthèse de toutes nos contributions.

## 3.1 Données cliniques et entrepôts de données

Les données produites dans le cadre du soin sont disponibles dans les systèmes d'informations hospitaliers (SIH) et peuvent *a priori* être exploitées à des fins de recherche. C'est le cas par exemple des études rétrospectives où des cliniciens peuvent recueillir et analyser les données passées d'une population spécifique qu'ils ont pris en charge. Les études rétrospectives reposent sur l'identification de la population de l'étude, ou *screening*, suivi de l'extraction d'informations structurées pour cette population (âge, sexe, valeurs biologiques ou codes issus du PMSI). Ces deux étapes peuvent être réalisées par les cliniciens directement dans le SIH. Le *screening* peut être réalisé par interrogation d'un système d'information local comme les codes PMSI *via* les départements d'informations médicales (DIM), ou de manière anticipée grâce à des listes de patients par pathologie tenues par les soignants. De la même manière, l'extraction des informations structurées pour cette population peut se faire par interrogation dans les bases de biologie ou dans les comptes rendus textuels lorsque ces informations ne sont pas disponibles dans les données structurées. Toutes ces démarches sont limitées par le temps humain nécessaire, notamment lorsque l'étude inclut des patients de différents services, et par la qualité des données disponibles dans le SIH.

Ces limitations sont progressivement surmontées grâce à la mise en place des EDS. Les bases de données que constituent les EDS concentrent différents flux d'information du SIH (données du PMSI, comptes rendus, données biologiques, etc.) avec parfois une nomenclature et un outil d'interrogation de la base communs à plusieurs centres hospitaliers. Selon les finalités déclarées pour l'EDS, les données à caractère personnel pouvant être incluses dans l'entrepôt peuvent varier.

Au CHU de Nantes, la solution déployée est eHOP (Madec et al., 2019), développée par l'équipe DOMASIA (DOnnées MAssives et Systèmes d'Information Apprenants en santé) du Laboratoire du Traitement du Signal et de l'Image de Rennes (LTSI - UMR 1099), dirigée par le Pr Marc Cuggia. Bien que l'outil se base sur des librairies *OpenSource*, eHOP est un outil commercialisé par la société ENOVACOM, filiale d'Orange Business Services. L'outil eHOP est déployé au sein du groupe HUGO (Hôpitaux Universitaires du Grand

Ouest) qui regroupe les CHU d'Angers, Brest, Nantes, Rennes et Tours ainsi que l'Institut de Cancérologie de l'Ouest. Il fournit différents connecteurs permettant d'intégrer dans les EDS les données hétérogènes issues des SIH.

Pour le CHU de Nantes, les flux de données intégrés dans l'EDS sont nombreux et en constante évolution. Parmi ces flux, on retrouve :

- un flux administratif permettant d'identifier les patients et de faire le lien avec les autres informations ;
- un flux de données de biologie ;
- un flux de comptes rendus (consultation, hospitalisation, opératoires, lettres de sortie, etc.) ;
- des flux issus du PMSI intégrant :
  - les codes de diagnostics CIM-10 associés au séjour. La Classification internationale des maladies (CIM), qui est recommandée par l'OMS, constitue le système de codage standard en matière de morbi-mortalité.
  - les actes médicaux (classification commune des actes médicaux ou CCAM). Cette classification permet de standardiser les actes médicaux réalisés par les professionnels de santé (actes diagnostiques, thérapeutiques, de suivi ou de prévention).
  - et les codes GHM (Groupe Homogène de Maladies). Le codage GHM permet de regrouper les patients selon des critères médicaux et diagnostiques similaires. Il est principalement utilisé à des fins administratives, notamment pour la gestion des ressources et le remboursement des soins par l'Assurance Maladie ;
- ou encore un flux médicament avec les prescriptions et les administrations de médicaments.

Ces flux peuvent être récupérés dans différents logiciels du SIH. Pour chaque logiciel, chaque flux de données aura son connecteur spécifique permettant d'alimenter l'EDS à un rythme régulier (quotidien, hebdomadaire ou mensuel selon le flux).

Une fois intégrées dans l'EDS, ces données peuvent être interrogées à travers un schéma de base de données relationnel Oracle *via* différentes applications :

- *via* l'interface eHOP, un requêteur avec une interface graphique « clique-boutons »,

- *via* tout moyen permettant une connexion à une base Oracle, par exemple avec le logiciel *Oracle SQL Developer* ou les langages de programmation R ou Python.

L'accès aux informations stockées dans un EDS impose certaines contraintes de confidentialité. L'usage des données directement identifiantes est restreint à la prise de contact avec les patients, soit pour leur proposer de participer à des études, soit à la suite de découvertes de caractéristiques génétiques ou de facteurs de risques à même de modifier leur prise en charge. Pour les autres finalités, seules des données désidentifiées ou pseudonymisées peuvent être employées. Comme stipulé dans l'exigence SEC-EXP-1 du référentiel CNIL sur les entrepôts de données, la manipulation de ces données désidentifiées ne peut se faire que dans des espaces de travail internes à l'entrepôt. Pour pouvoir travailler en dehors de ces espaces de travail ou partager des données issues de l'EDS, un processus d'anonymisation conforme aux trois critères de l'article du G29 n°05/2014 sur la protection des données doit avoir été appliqué aux données. L'exigence SEC-EXP-1 du référentiel indique que ce processus doit être documenté et démontrable. Si ces trois critères ne peuvent être réunis, une étude des risques de ré-identification doit être menée et documentée.

Cette exigence s'applique à la fois aux données brutes directement extraites de l'EDS comme aux données transformées à partir des données brutes, par exemple les algorithmes d'intelligence artificielle appris à partir des données brutes. Si des solutions existent pour anonymiser un jeu de données tabulaires (Guillaudeux et al., 2023), il est plus difficile de démontrer le caractère anonyme des données textuelles ou encore des modèles de langues pré-entraînés à partir des données textuelles.

Ces contraintes liées au caractère personnel des données poussent à faire des choix techniques sur différents aspects d'un projet de TAL afin de respecter ce cadre réglementaire : choix des outils d'annotation, circuit des données entre les différents partenaires, choix des algorithmes selon les ressources de calcul disponibles en interne, etc.

Les deux corpus présentés ensuite dans ce chapitre sont composés de données issues de l'EDS du CHU de Nantes.

## 3.2 Projet GAVROCHE

Le projet GAVROCHE est l'un des projets lauréats de l'Appel à Projet inter-régional des entrepôts du groupe HUGO (AAP-GIRCI Grand Ouest) en septembre 2019. GAVROCHE est un projet multicentrique inter-régional exploitant les entrepôts de données de santé (EDS) du grand ouest, soit cinq CHU - Angers, Brest, Nantes, Rennes et Tours. Le projet est porté par le Pr Hadjadj et le Dr Wargny du CHU de Nantes avec le soutien de la Clinique des Données.

Ce projet s'intéresse, au travers d'une étude rétrospective, à l'association entre la variabilité glycémique des premiers jours suivant l'admission et le décès, chez des patients hospitalisés pour insuffisance cardiaque aiguë (ICA). L'insuffisance cardiaque (IC) est l'incapacité du cœur à assurer un débit permettant l'ensemble des fonctions normales de l'organisme (Denolin et al., 1983). La prévalence de l'IC augmente et est en passe de devenir la complication cardiovasculaire la plus fréquente. C'est une pathologie à la fois fréquente et grave comme le souligne la mortalité importante associée à un épisode d'insuffisance cardiaque aiguë (ICA) (Arrigo et al., 2020).

D'ordinaire, les études rétrospectives s'appuient sur des données disponibles à large échelle et directement exploitables à des fins de recherche comme les données structurées renseignées dans les systèmes d'informations hospitaliers (SIH) (poids, taille, IMC, âge, sexe...), les données de biologie ou encore les données du PMSI, comme les actes CCAM nous donnant l'information qu'un patient a bien bénéficié de tel examen ou de tel geste chirurgical.

D'autres données, comme les comptes rendus hospitaliers (CRH), rassemblent des informations avec un degré de granularité plus fin. Ils sont cependant moins accessibles à large échelle car ils nécessitent une lecture attentive pour en extraire les informations d'intérêt. Dans le cas d'une étude sur une maladie rare incluant quelques dizaines de patients, les informations d'intérêt sont collectées manuellement par un humain dans chaque compte rendu. Cependant, lorsqu'une étude compte plusieurs milliers de patients, ce recueil manuel des informations n'est plus possible et il faut automatiser cette collecte de données en s'appuyant sur des méthodes du traitement automatique des langues.

Dans le projet GAVROCHE, la majorité des données brutes (c'est-à-dire les données structurées et non structurées telles qu'elles sont disponibles dans l'EDS) est accessible dans l'EDS et requêtable dans tous les centres de l'inter-région grâce à un schéma relationnel très proche et au logiciel commun eHOP. Bien que ces données soient disponibles,

elles ne sont pas directement exploitables dans leur forme brute. Les données structurées nécessitent d'être uniformisées, notamment entre les centres où elles peuvent différer. Quant aux CRH, leur exploitation requiert l'usage de méthodes issues du TAL.

Les objectifs de l'étude GAVROCHE sont multiples. Pour la recherche en épidémiologie clinique, l'objectif principal est l'analyse du rôle pronostique de la variabilité glycémique sur la mortalité associée à l'ICA chez les sujets présentant ou non un diabète. Un objectif secondaire de l'étude est l'analyse du rôle pronostique de la glycémie à l'admission sur la mortalité. Pour la recherche en données massives en santé, la réalisation du projet se fonde sur deux hypothèses : (i) la possibilité d'extraction d'informations médicales structurées à partir des CRH et (ii) l'accès aux données et la gouvernance au sein du Réseau inter-régional des Centres de Données Cliniques (RiCDC).

Nous présentons les méthodes de TAL mises en place dans le but d'extraire des informations médicales structurées à partir de CRH. Nous donnons les éléments de contexte nécessaires à la compréhension du projet, sans rentrer dans les détails portant sur les aspects généraux et administratifs liés au contexte inter-régional, comme le circuit global des données ou encore les aspects éthiques et réglementaires. Dans un premier temps, nous dressons un tableau d'ensemble du travail de TAL réalisé dans ce projet en listant les points clés de sa réalisation ainsi que les contraintes techniques liées à sa dimension inter-régionale. Toutes les étapes en amont de l'annotation sont décrites : la constitution du corpus, le choix de l'outil d'annotation et la rédaction du guide d'annotation. Nous présentons également le processus d'annotation de deux tâches : la classification de texte pour vérifier si les séjours sont bien associés à une ICA, et la reconnaissance d'entités nommées pour extraire des informations structurées à partir des CRH. Enfin, nous décrivons les phases d'adjudication entre les cliniciens et clôturons cette partie en donnant quelques statistiques sur le corpus annoté qui sera produit.

### 3.2.1 Vue d'ensemble du TAL dans GAVROCHE

Le travail de TAL nécessaire à GAVROCHE peut être résumé en trois grandes étapes successives. La première est la phase d'apprentissage de l'algorithme pour chaque tâche à savoir la classification de textes et la reconnaissance d'entités nommées. Cette phase est monocentrique basée uniquement sur les données du CHU de Nantes. La deuxième étape est la validation des algorithmes sur un échantillon de CRH annotés au niveau de chaque centre impliqué dans le projet. Enfin, la dernière étape vise à déployer les algorithmes à grande échelle sur tous les CRH, et ce, dans tous les centres. Chacune de ces étapes est

détaillée dans le plan ci-dessous.

**(1) La phase d'apprentissage des algorithmes, monocentrique, basée uniquement sur les CRH annotés au CHU de Nantes**

- Les CRH des séjours sont tout d'abord extraits de l'entrepôt de données du CHU de Nantes selon les critères d'éligibilité définis pour le projet.
- De l'ensemble des CRH extraits, un échantillon est sélectionné pour réaliser la première phase d'annotation, effectuée pour la tâche de classification de texte qui vise à vérifier si les CRH sont bien associés à un séjour pour ICA.
- Une fois la première phase d'annotation terminée, les CRH identifiés comme associés à un séjour pour ICA sont annotés de nouveau par les experts nantais, cette fois-ci pour la tâche de reconnaissance d'entités nommées afin d'identifier une série de variables prédéfinies.
- Pour chacune des deux phases d'annotation :
  - un premier échantillon de CRH annotés est utilisé pour l'apprentissage de l'algorithme ;
  - le reste de l'échantillon sert de base de validation pour vérifier la fiabilité de l'algorithme.

**(2) La phase de validation des algorithmes réalisée indépendamment dans chaque centre**

- Les premières étapes de la phase de validation sont identiques à celles présentées dans la phase d'apprentissage au CHU de Nantes
  - Les CRH des séjours sont extraits de l'entrepôt de données local selon les mêmes critères d'éligibilité.
  - Les deux phases d'annotations réalisées dans la phase d'apprentissage sont reproduites à l'identique par des experts locaux, à la seule différence que l'échantillon de CRH est moins conséquent puisqu'il sert uniquement à vérifier la fiabilité de l'algorithme.

- En parallèle, les algorithmes développés pour chaque tâche dans la première phase sont appliqués aux mêmes échantillons de CRH.
- Les résultats des algorithmes sont ensuite confrontés avec les annotations des experts et les discordances sont analysées pour déterminer qui de l'algorithme ou des experts a raison.
- À l'issue de cette phase, nous avons les performances de l'algorithme pour chaque variable dans tous les centres. Selon les performances, nous pourrons décider pour chaque variable (i) d'abandonner cette variable si elle n'est pas suffisamment présente dans les CRH, (ii) de revenir à la phase 1 pour annoter davantage de CRH si la variable est présente, mais pas en quantité suffisante ou (iii) considérer que l'algorithme est valide si les performances sont suffisantes.

**(3) L'application des algorithmes à grande échelle**

- L'algorithme de classification de texte est appliqué, pour chaque centre, à tous les CRH récupérés par la procédure SQL. Tous les séjours ayant au moins un CRH classé comme associé à une ICA sont conservés.
- L'algorithme d'extraction d'informations est appliqué aux CRH sélectionnés par l'algorithme de classification de texte, pour chaque variable validée et chaque centre, afin de produire les données finales nécessaires au projet.

**Contraintes de mise en œuvre du TAL**

Le contexte inter-régional du projet ajoute des contraintes techniques non négligeables à la mise en place du TAL pouvant compromettre sa réalisation à l'échelle régionale. En particulier, le partage de données entre les centres est limité dû à la confidentialité des patients. Il faut alors trouver des solutions pour limiter le partage de données entre les centres. Il n'est pas question ici d'apprentissage fédéré. Dans le cas de GAVROCHE, le choix a été fait de partager les algorithmes plutôt que les données. Plus précisément, les algorithmes sont développés uniquement sur le corpus annoté au CHU de Nantes puis envoyés aux autres centres pour être appliqués sur les corpus annotés localement. Avec cette approche, nous n'avons pas accès aux CRH annotés par les experts des autres centres et cela implique d'appliquer les algorithmes « à l'aveugle » sur les données. Comme la qualité des données varie d'un centre à l'autre, nous n'avons pas la garantie que l'exécution

des algorithmes se déroule dans les mêmes conditions que sur le corpus nantais, ni que les performances des algorithmes soient équivalentes.

Une seconde contrainte porte sur la disponibilité de ressources humaines. En effet, dans les autres centres, nos algorithmes sont appliqués par les experts locaux sur le corpus annoté auquel ils ont accès. Nous sommes donc dépendants de la disponibilité et des compétences techniques des experts locaux pour prendre en main nos algorithmes. Une documentation décrivant la procédure à suivre est envoyé dans chaque centre pour les guider dans l'usage des algorithmes, mais cela reste tout de même plus chronophage et fastidieux que d'accéder nous-mêmes aux données et réaliser toutes les étapes.

Une dernière contrainte est l'accès aux ressources de calcul dont les établissements de santé sont rarement dotés pour la recherche. Dans une ère où le TAL est principalement mis en œuvre à l'aide d'approches statistiques nécessitant des ressources de calcul importantes, c'est une contrainte bloquante. Nous anticipons le fait que nous n'exploiterons pas tout le potentiel des méthodes état de l'art. Nous serons donc limités dans le choix des approches qui seront déployées dans les centres de l'inter-région et devrons miser sur des algorithmes moins gourmands pouvant être exécutés sur des machines quelconques. Nous nous attendons à ce que cette contrainte ait un impact non négligeable sur les résultats finaux obtenus par le TAL et plus globalement sur l'étude finale du projet GAVROCHE.

### 3.2.2 Préparation de l'annotation

Plusieurs éléments sont nécessaires en amont de l'annotation d'un corpus. Il s'agira tout d'abord de constituer le corpus brut qui fera l'objet de l'annotation. Selon la nature du corpus, des prétraitements sont nécessaires pour homogénéiser les données. La tâche d'annotation doit ensuite être définie ainsi que l'outil qui permettra de la réaliser. L'annotation est un processus itératif, les décisions prises au début du projet peuvent évoluer tout au long du processus grâce aux retours d'expérience des annotateurs. Ces retours d'expériences peuvent nous amener à faire des modifications à la fois sur l'outil d'annotation, mais aussi sur les questions scientifiques, en ajustant les définitions des informations à annoter après avoir été au contact des données. Nous décrivons ici toutes les étapes que nous avons réalisées pour l'annotation du corpus pour le projet GAVROCHE.

#### 3.2.2.1 Constitution du corpus et prétraitements

La population d'étude est composée de patients hospitalisés pour insuffisance cardiaque aiguë (ICA) sévère entre 2011 et 2019. L'unité statistique principale dans cette étude est le séjour et non le patient. Nous pouvons donc avoir plusieurs séjours d'intérêt pour un même patient.

Les critères d'inclusion des séjours d'intérêt dans l'étude sont les suivants :

- Femme ou homme d'âge $\geq$ 18 ans.
- Hospitalisé(e) dans l'un des centres hospitaliers du groupe HUGO (Angers, Brest, Nantes, Rennes et Tours) entre le 1er janvier 2011 et le 31 décembre 2019 dans la limite des données disponibles dans les EDS de chaque centre.
- Dont le séjour hospitalier est associé, dans le PMSI, à un code GHM (Groupes Homogènes de Malades) commençant par « 05M09 » (insuffisance cardiaque - état de choc circulatoire) et/ou à un code CIM-10 d'insuffisance cardiaque (I50, I11.0, I13.0 I13.2, I13.9 et R570) en diagnostic principal ou relié.
- Ayant au moins un CRH, une lettre de sortie ou une lettre de synthèse associée au séjour dans l'EDS (ou tout document jugé équivalent par le personnel du CDC local).
- Ayant au moins une valeur de glycémie disponible dans les 24 heures suivant l'admission. Ce critère est nécessaire pour l'objectif secondaire même si trois glycémies sont requises pour le calcul de la variabilité glycémique et donc pour l'objectif principal.
- Personnes dont le traitement des données à des fins de recherche dans l'EDS local a été autorisé dans les limites approuvées par la CNIL.

Les critères de non-inclusion sont les suivants :

- Séjour sans nuitée et non soldé en raison du décès du patient (remarque : ajout postérieur à la soumission CNIL afin d'exclure les hospitalisations de jour, non identifiables aisément dans le SI).
- Personne ayant manifesté sa volonté de ne pas voir ses données utilisées à des fins de recherche par l'établissement.

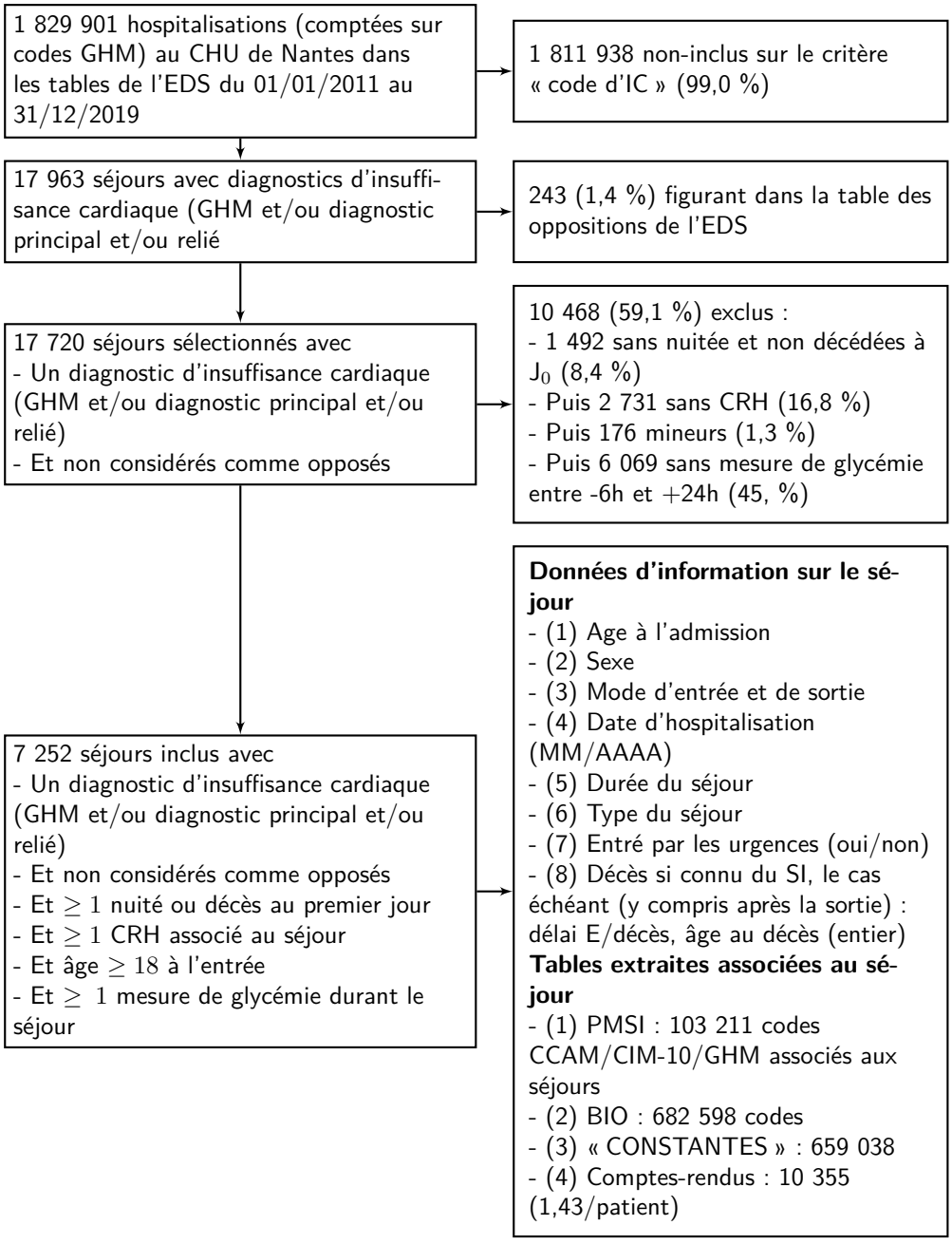


FIGURE 3.1 : Flow-chart de l'identification des séjours d'intérêts pour GAVROCHE, obtenu par procédure SQL. Schéma des flux de données repris de la thèse du Dr Matthieu Wargny (2022).

Les CRH bruts des 7 252 séjours d'intérêt sont récupérés dans l'entrepôt de données du CHU de Nantes par une procédure SQL. Le schéma résumant le flux de données associé est visible en figure 3.1.

Les CRH sont des textes rédigés sans contraintes. La structure et la qualité de ces documents sont hétérogènes, car rédigés par différents cliniciens provenant de différents services, avec des outils métiers de traitement de textes pouvant changer au cours du temps. Cette hétérogénéité des documents implique la mise en place de prétraitements pour les homogénéiser et assurer la continuité et la fluidité entre les différents outils utilisés pour le TAL, que ce soit l'outil d'annotation ou les algorithmes d'extraction d'informations. Les prétraitements sur les CRH sont effectuées avant leur annotation.

Bien que le contenu même des CRH ne suive aucune structure de rédaction, un même modèle est partagé par les documents issus de l'EDS. Ce modèle est composé de trois sections distinctes :

- L'en-tête : avec des informations propres au service qui a généré ce compte rendu, le début de l'en-tête est identique pour tous les documents de même provenance. L'en-tête contient également des informations administratives déjà présentes dans les données structurées de l'entrepôt de données.
- Le corps du compte rendu : avec des informations relatives à la santé du patient faisant l'objet du compte rendu.
- Le pied de page : avec des informations propres au service dont provient le document, à l'instar de l'en-tête.

Le premier traitement appliqué aux CRH est la segmentation des CRH en ces trois sections. Comme l'en-tête et le pied de page des comptes rendus ne contiennent pas d'informations médicales utiles pour l'annotation et occupent une grande partie des documents, nous procédons à un prétraitement pour les retirer. En parcourant visuellement un échantillon de comptes-rendus, des marqueurs récurrents ont été identifiés, permettant ainsi de séparer automatiquement ces trois parties à l'aide d'un algorithme à base d'expressions régulières. En effet, le corps des comptes rendus commence quasi systématiquement par une formule de politesse, telle que « Cher Confrère », « Chère Consœur », « Madame, » ou encore « Monsieur, ». Ce premier indicateur permet alors de séparer l'en-tête du corps du compte rendu. À l'instar du corps du compte rendu, un marqueur, « .SMSDESTINATAIRES », apparaît de façon systématique au début des pieds de pages et permet ainsi de séparer le corps du compte rendu du pied de page.

La segmentation en mots est certainement le prétraitement le plus important à appliquer puisque tous les prétraitements ultérieurs dépendent de celui-ci. Néanmoins, la nature biomédicale des documents rend ce traitement complexe et délicat (Cruz Díaz and Maña López, 2015). En effet, les marqueurs classiques de ponctuation peuvent être omis ou insérés de manière impromptue dans les CRH, certains termes employés tels que des noms d'organismes, des valeurs numériques ou encore des termes spécifiques au domaine médical peuvent également contenir des marques de ponctuation. Pour faciliter l'application des algorithmes de segmentation, les prétraitements suivants ont été appliqués aux CRH :

- suppression des balises propres aux CRH, tels que « .smsdebut » ou « .smsfin ».
- suppression de la ponctuation dans les numéros de téléphone.
- normalisation des acronymes (exemple : « A. I. N. S. » devient « A.I.N.S. »).

#### 3.2.2.2 Sélection de l'outil d'annotation

Les outils d'annotation sont essentiels pour le développement et l'évaluation des algorithmes de TAL. De nombreux outils d'annotation existent et proposent différentes fonctionnalités afin de s'adapter aux multiples besoins (Neves and Ševa, 2021). La sélection d'un outil d'annotation repose sur plusieurs critères, ceux-ci étant définis par la tâche visée et la nature des données à annoter. Dans le cas de GAVROCHE, les critères suivants étaient à prendre en compte :

- La sécurité des données. En effet, étant donné la nature sensible des données faisant l'objet de cette annotation, les outils nécessitant un téléversement des comptes rendus sur une plateforme externe au CHU étaient d'emblée exclus.
- Le public visé par l'annotation manuelle. Dans le cadre de ce projet, les annotateurs sont des cliniciens qui effectuent cette tâche d'annotation sur leur temps de travail. Il faut par conséquent que l'outil soit simple, efficace et qu'il ne nécessite pas d'installation côté utilisateur.

À partir de ces critères, plusieurs outils d'annotation ont été testés : Oxygen XML, TXM (Heiden et al., 2010) et Prodigy (Montani and Honnibal, 2018). Après plusieurs mois de tests avec les annotateurs, l'outil Prodigy a été retenu comme solution. Prodigy a été installé sur un serveur CHU accessible depuis l'intranet par simple connexion à

un navigateur Internet (Google Chrome, Mozilla Firefox notamment) et a pu être testé dans un premier temps grâce à une licence de recherche fournie gratuitement par les développeurs du logiciel. Une fois l'outil retenu pour le projet, cinq licences ont été achetées (une licence distribuée aux cinq centres participants) pour le déploiement de l'outil. Ces licences n'ont pas de limites temporelles et peuvent être utilisées par plusieurs utilisateurs, permettant la réutilisation de l'outil dans d'autres projets des CHU de l'inter-région.

Prodigy propose une interface minimaliste et ergonomique qui a su séduire les annotateurs, notamment grâce aux raccourcis clavier tels que la sélection des étiquettes d'annotation avec le pavé numérique ou le passage au document suivant par pression sur la touche entrée. Le seul inconvénient auquel les annotateurs ont dû s'adapter est l'impossibilité de revenir sur un document sauvegardé après l'annotation. En effet, Prodigy garde dans un historique les 15 derniers documents annotés non sauvegardés, mais une fois sauvegardés, ces documents sont retirés de l'historique et envoyés dans la base de données sqlite et ne peuvent plus être modifiés par l'interface.

Contrairement à d'autres outils d'annotation, comme BRAT (Stenetorp et al., 2012), où l'annotation se fait au niveau caractères, l'annotation des documents dans Prodigy est réalisée au niveau mot. Prodigy étant un logiciel développé par Explosion, éditeur logiciel qui développe également la librairie spacy, tous les prétraitements de documents dont la segmentation en mots sont réalisés nativement dans Prodigy avec spacy. Les formats de données utilisés dans l'outil sont également standards et permettent d'être réutilisés facilement. Les documents en entrée sont stockés sous forme de textes libres dans un document json avec leurs métadonnées associées : identifiant du document et identifiant du séjour patient. Lors de l'annotation, les documents et leurs annotations sont stockées dans une base de données sqlite. Une fois l'annotation terminée, les CRH et leurs annotations peuvent être extraits au format json à partir de la base sqlite. Pour chaque document annoté, on retrouve les éléments suivants dans le json :

- « text » : le texte original au format libre,
- « tokens » : le texte original segmenté en mots,
- « meta » : les métadonnées associées au document,
- « spans » : la liste de toutes les annotations réalisées dans le document. Chaque annotation contient les éléments suivants :
  - les index de caractères de l'annotation,

– les index de mots de l'annotation,

– l'étiquette associée à la portion de texte annotée.

#### 3.2.2.3 Définition des tâches

L'annotation des comptes rendus hospitaliers s'est déroulée en deux étapes, à travers deux tâches. La première tâche de classification de texte a pour objet de vérifier la pertinence de l'inclusion des séjours basée sur les codes GHM et CIM-10. La seconde tâche est la reconnaissance d'entités nommées à travers 45 étiquettes correspondant à 24 variables cliniques. Les deux tâches sont décrites ci-dessous et formalisées par un guide d'annotation.

**Classification de texte - Validation des séjours pour ICA**

La tâche de classification de textes est effectuée en premier. Elle permet de valider la sélection des CRH extraits de l'EDS selon critères d'inclusion de la figure 3.1 et d'exclure les faux positifs, i.e. les CRH non associés à une hospitalisation pour ICA comme la chirurgie aortique ou greffe cardiaque programmée et la décompensation cardiaque secondaire mais non présente à l'admission. C'est une tâche idéale pour prendre en main l'outil d'annotation, car seuls trois choix sont possibles pour classer le CRH : « oui », « non », « incertain ». Même si l'objectif est d'avoir uniquement deux classes pour cette tâche (« oui » et « non »), nous avons introduit la classe « incertain » pour permettre aux cliniciens de discuter des cas difficiles. Les CRH annotés comme « incertain » seront ensuite reclassés, soit en cas valide, soit en cas invalide à la suite de l'adjudication réalisée par les cliniciens. En plus de la classification de texte, des annotations sous la forme d'étendues de texte ont été ajoutées pour justifier de la classification du document. Annoter cette tâche comme telle demande plus de temps de travail lors de l'annotation, mais fourni une garantie d'explicabilité de la classification du document. Cela facilite également la discussion au moment de l'adjudication en comparant les annotations sans avoir à relire intégralement le document, notamment lorsqu'il y a des discordances entre les annotateurs.

**Reconnaissance d'entités nommées - Annotation des variables d'intérêt**

La tâche de reconnaissance d'entités nommées est dédiée aux variables cliniques d'intérêt pour le projet. Elle intervient après la tâche de classification de texte où seuls les séjours pour ICA validés en amont sont annotés. Les variables cliniques d'intérêt, au nombre de

24, ont été définies par les cliniciens porteurs du projet. Chaque variable est annotée à travers un nombre d'étiquettes associées, allant de 1 à 7 étiquettes. Au total, 45 étiquettes ont été retenues pour l'annotation, réparties en 5 blocs d'annotation. Un bloc d'annotation correspond à un ensemble comprenant un maximum de 10 étiquettes à annoter à la lecture d'un CRH donné. À des fins d'ergonomie, nous avons rapidement et de façon consensuelle acté la nécessité de conserver une cohérence sémantique à chaque bloc, de ne pas dépasser 10 étiquettes pour permettre à l'annotateur d'utiliser uniquement les raccourcis sur le clavier numérique. Ce choix d'effectuer plusieurs lectures d'un même CRH pour annoter toutes les étiquettes peut sembler contre-productif au premier abord. Pourtant l'annotation par groupe d'étiquettes nous a permis d'accélérer l'annotation de 30 %, passant d'une moyenne de 10 minutes par CRH à 7 minutes pour annoter toutes les étiquettes. De plus, les annotations sont de meilleure qualité grâce à une charge cognitive des annotateurs fortement réduite. Les blocs d'annotation ont beaucoup évolué en 3 ans suite aux retours des différents cliniciens et à l'expérience même de l'annotation au contact des données partagée régulièrement avec l'équipe. La configuration des blocs a ensuite été figée afin de débuter l'annotation finale qui sera utilisée pour développer les systèmes de TAL.

Le bloc 1 rassemble les facteurs déclenchant de l'ICA (annotés à travers 7 étiquettes), la notion de premier épisode d'ICA du patient, le traitement habituel et la notion d'arrêt cardiaque à l'admission. Dans le bloc 2, le facteur étiologique de la cardiopathie causale est annoté à travers cinq étiquettes (ischémique, valvulaire, rythmique, autre et non connu), tandis que le type d'ICA est annoté à travers quatre étiquettes (IC droite isolée, IC gauche décompensée, OAP et choc cardiogénique). On retrouve dans le bloc 3 les antécédents respiratoires (insuffisance respiratoire chronique, bronchopneumopathie chronique obstructive et syndrome d'apnée du sommeil), d'AVC et de diabète. Les informations cliniques à l'admission, telles que la fréquence cardiaque, la pression artérielle, le poids, la taille et l'IMC, sont regroupées dans le bloc 4. On retrouve également le tabagisme dans ce bloc. Enfin, le bloc 5 rassemble l'AC/FA à l'admission, la valeur de la fraction d'éjection ventriculaire gauche et les antécédents de trouble du rythme, de dépression, de troubles cognitifs et de cancer. Un tableau récapitulatif est proposé en annexes 6.1 pour chacun des blocs.

### 3.2.3 Annotation pour la classification de textes : validation des séjours pour ICA

L'annotation de la tâche de classification de texte a été réalisée par trois annotateurs : Pr Samy Hadjadj (SH), Dr Matthieu Wargny (MW) et Dr Florian Martin (FM). Chaque annotateur avait un maximum 900 CRH à annoter selon leurs disponibilités, dont au moins 200 CRH en commun avec les autres annotateurs pour calculer les accords inter-annotateurs. Pour vérifier la qualité de la sélection des séjours avec les codes PMSI, les 900 CRH par annotateur sont distribués comme ceci :

- 300 CRH avec uniquement un code CIM-10,
- 300 CRH avec uniquement un code GHM,
- 300 CRH avec un code CIM-10 et un code GHM.

Au total, 1 639 CRH uniques ont été annotés, comme présenté dans le tableau 3.1. Sur cet échantillon, 1 063 CRH (64,86 %) correspondent bien à des séjours pour ICA, 482 CRH (29,40 %) comme ne représentant pas une ICA et 94 CRH (5,74 %) ont été annotés comme incertain et seront rediscutés entre les cliniciens.

| Annotateur | Nb. CRH annotés | ICA oui | ICA non | ICA incertain |
|---|---:|---:|---:|---:|
| MW | 898 | 590 | 251 | 57 |
| SH | 162 | 122 | 37 | 3 |
| FM | 898 | 541 | 298 | 59 |
| Nb. CRH uniques | 1 639 | 1 063 | 482 | 94 |

Tableau 3.1 : Distribution du corpus annoté pour la tâche de classification de texte.

#### 3.2.3.1 Mesures d'accord inter-annotateurs pour la classification de textes

À partir des annotations réalisées, les accords inter-annotateurs ont été mesurés avec quatre indicateurs : (1) proportion d'accord global (ou pourcentage d'accord), (2) Kappa de Cohen, (3) Gwet AC1, (4) Kappa de Fleiss.

**Pourcentage d'accord** La proportion d'accord global $P_o$ est formulé comme suit :

$$P_o = \frac{\text{Nombre d'accords observés}}{\text{Nombre total de documents annotés}}$$

**Kappa de Cohen** (Cohen, 1960) mesure l'accord entre deux annotateurs au-delà de l'accord qui serait obtenu par hasard. Cette mesure se base sur la comparaison des annotations réelles avec les annotations qui seraient attendues si les annotateurs agissaient uniquement par hasard. Le Kappa de Cohen est formulé comme suit :

$$\kappa_{Cohen} = \frac{P_o - P_e}{1 - P_e}$$

où $P_o$ est la proportion d'accord observée, ou pourcentage d'accord, entre deux annotateurs et $P_e$ est la probabilité d'un accord aléatoire.

Dans le Kappa de Cohen, la probabilité d'un accord aléatoire, $P_e$, est estimé par la probabilité qu'une classe soit utilisée par un annotateur :

$$P_e(\kappa_{Cohen}) = \frac{1}{N^2} \sum_{q=1}^{K} f_q \cdot g_q$$

où $K$ correspond au nombre total de classes et $q$ à une classe spécifique. $f_q$ et $g_q$ correspondent au nombre de fois où la classe $q$ a été utilisée par les annotateurs $f$ et $g$. $N$ correspond au nombre total de documents annotés.

Dans le cas d'une classification binaire ($K = 2$), $P_e$ serait estimé comme suit, où $f_1$, $g_1$, $f_2$ et $g_2$ sont définis à l'aide du tableau 3.2 :

$$P_e(\kappa_{Cohen}) = \frac{f_1 \cdot g_1 + f_2 \cdot g_2}{N^2}$$

| | Annotateur B | | |
|---|---|---|---|
| Annotateur A | 1 | 2 | Total |
| 1 | $a$ | $b$ | $g_1 = a + b$ |
| 2 | $c$ | $d$ | $g_2 = c + d$ |
| Total | $f_1 = a + c$ | $f_2 = b + d$ | $N$ |

Tableau 3.2 : Exemple de tableau permettant de calculer les accords inter-annotateurs pour une tâche de classification binaire.

**Gwet AC1** (Gwet, 2014) est une mesure d'accord entre deux annotateurs, semblable au Kappa de Cohen. Le Gwet AC1 est formulé comme le Kappa de Cohen, la seule différence réside dans la manière dont est estimée la probabilité d'un accord aléatoire $P_e$. En effet, contrairement au Kappa de Cohen où l'accord aléatoire est fondé sur la probabilité que les

deux annotateurs classent par hasard un même document avec la même classe, Gwet AC1 est fondé sur la probabilité qu'au moins un des annotateurs classe aléatoirement et que seule une proportion inconnue des annotations est aléatoire. De plus, Gwet AC1 prend en compte la prévalence des classes dans son estimation de l'accord aléatoire, ce qui lui permet d'être plus robuste dans les scénarios où la distribution entre les classes n'est pas équilibrée (Gwet, 2001). Le Gwet AC1 est formulé comme suit :

$$AC1 = \frac{P_o - P_e}{1 - P_e}$$

où $P_o$ est la proportion d'accord observée, ou pourcentage d'accord, entre deux annotateurs et $P_e$ est la probabilité d'un accord aléatoire, calculée comme suit :

$$P_e(AC1) = \frac{1}{K-1} \sum_{q=1}^{K} \left( \frac{N_q}{N} \cdot \frac{N - N_q}{N} \right)$$

où $K$ correspond au nombre total de classes et $q$ à une classe spécifique. $N_q$ correspond au nombre moyen de fois qu'une certaine classe est utilisée par un annotateur. Dans le cas de deux annotateurs, $N_q$ vaut $(f_q + g_q)/2$. $\frac{N_q}{N}$ représente le pourcentage de documents associés à la classe $q$ et $\frac{N-N_q}{N}$ représente le pourcentage de documents associés à une classe autre que $q$.

Dans le cas d'une classification binaire ($K = 2$), $P_e$ serait estimé comme suit à l'aide du tableau 3.2 :

$$P_e(AC1) = \left( \frac{(f_1 + g_1)}{2N} \cdot \frac{N - \frac{(f_1+g_1)}{2}}{N} \right) + \left( \frac{(f_2 + g_2)}{2N} \cdot \frac{N - \frac{(f_2+g_2)}{2}}{N} \right)$$

Ces différences de formulation de l'accord aléatoire entre le Kappa de Cohen et le Gwet AC1 amènent à des résultats drastiquement différents. L'accord aléatoire varie entre 0 et 1 dans le Kappa de Cohen alors qu'il varie entre 0 et 0,5 dans le Gwet AC1. Cette limite à 0,5 du Gwet AC1 permet d'éviter les instabilités à l'origine du paradoxe de Kappa (Feinstein and Cicchetti, 1990) - c'est-à-dire avoir un Kappa faible malgré un accord élevé entre les annotateurs.

Plusieurs travaux ont comparé en détails les différences entre le Kappa de Cohen et le Gwet AC1 à travers divers scénarios et ont montré que le Gwet AC1 est une mesure plus robuste que le Kappa de Cohen (Wongpakaran et al., 2013; Hoek and Scholman, 2017).

**Kappa de Fleiss** (Fleiss, 1971) est une généralisation du Pi de Scott (1955) à un nombre d'annotateurs supérieur à deux. Le Kappa de Fleiss est formulé comme suit :

$$\kappa_{Fleiss} = \frac{P_o - P_e}{1 - P_e}$$

où $P_o$ est la proportion d'accord observée entre tous les annotateurs et $P_e$ est la probabilité d'un accord aléatoire.

$$P_o(\kappa_{Fleiss}) = \frac{1}{Nn(n-1)} \left( \left( \sum_{i=1}^{N} \sum_{q=1}^{K} n_{iq}^2 \right) - Nn \right)$$

$$P_e(\kappa_{Fleiss}) = \sum_{q=1}^{K} \left( \frac{1}{Nn} \sum_{i=1}^{N} n_{iq} \right)^2$$

où $N$ est le nombre total de documents et $i$ un document spécifique. $n$ est le nombre total d'annotateurs. $K$ est le nombre total de classes possibles et $q$ une classe spécifique. $n_{iq}$ représente le nombre d'annotateurs qui ont attribué la classe $q$ au document $i$.

#### 3.2.3.2 Résultats des mesures d'accord inter-annotateurs et adjudication

Les accords inter-annotateurs mesurés sont présentés dans le tableau 3.3. D'après le tableau d'interprétation du Kappa de Cohen proposé par Landis and Koch (1977), un accord moyen d'environ 0,60 correspond à un accord modéré et proche d'un accord fort.

| | Nb. de CRH annotés | Kappa Fleiss | Kappa Cohen | Gwet AC1 | Accord global |
|---|---|---|---|---|---|
| 3 annotateurs | 41 | 0,64 | - | - | 0,80 |
| MW & SH | 56 | - | 0,57 | 0,75 | 0,84 |
| MW & FM | 205 | - | 0,63 | 0,70 | 0,83 |
| SH & FM | 74 | - | 0,60 | 0,695 | 0,82 |

Tableau 3.3 : Accord inter-annotateurs globaux pour la tâche de classification de texte.

Après avoir calculé les accords inter-annotateurs, les cliniciens se sont réunis pour une séance d'adjudication afin de discuter des discordances, notamment pour les documents dont le cas d'ICA était incertain. Ces cas cliniques étaient difficiles à adjudiquer par les cliniciens experts. Par exemple, les cas ne présentant aucun signe de défaillance cardiaque autre qu'une insuffisance rénale chronique, et où aucune affirmation d'insuffisance car-

diaque n'était présente dans le CRH. Suite à cette adjudication, les cas incertains ont été annotés comme n'étant pas des cas d'ICA et nous obtenons la distribution présentée dans le tableau 3.4 pour les deux classes finales.

| | ICA validés | ICA non validés | Total |
|---|---|---|---|
| Nombre de CRH | 1 063 | 576 | 1 639 |
| Nombre de phrases | 61 219 | 34 561 | 95 780 |
| Nombre de mots | 843 755 | 439 199 | 1 282 954 |
| Nombre de mots moyen par CRH | 793 | 762 | 783 |
| Nombre de mots min. | 78 | 45 | 45 |
| Nombre de mots max. | 4 142 | 3 252 | 4 142 |

TABLEAU 3.4 : Description du jeu de données final après adjudication pour la tâche de classification de texte.

### 3.2.4 Annotation pour la reconnaissance d'entités nommées : extraction des variables d'intérêt

Une fois la liste des étiquettes établie et leur répartition dans les blocs 1 à 5 figées, l'annotation a pu débuter. L'annotation du corpus s'est déroulée en deux temps. Dans un premier temps, une annotation commune d'un ensemble de CRH a été réalisée par le groupe d'annotateurs en vue de mesurer les accords inter-annotateurs et d'éprouver la qualité du guide d'annotation. Pour ce faire, 50 CRH ont été annotés par 7 annotateurs, dont trois endocrinologues (S. Smati, P. Morcel, E. Scharbarg et S. Hadjadj), un médecin neurovasculaire (P. Constant dit Beaufils), un cardiologue (D. Stévant) et un médecin de santé publique (M. Wargny). Un supplément de 50 CRH a été proposé aux 3 annotateurs les plus aguerris (PCDB, SH et MW) afin d'améliorer le guide d'annotation. À l'issue de cette première phase d'annotation, les annotations réalisées ont été confrontées et une phase d'adjudication entre les annotateurs a été organisé pour statuer sur les discordances.

Les annotateurs ont ensuite été sollicités de nouveau pour seconde phase d'annotation. Cette fois-ci, il n'y a plus de documents en commun entre les annotateurs, chaque CRH est annoté par un seul annotateur. Un total de 400 CRH supplémentaires ont été distribués parmi 5 annotateurs, chaque annotateur ayant entre 50 et 100 CRH selon leurs disponibilités. En regroupant l'ensemble des documents annotés dans ces deux phases d'annotation, nous arrivons à un total de 500 CRH annotés.

Avant de présenter en détails ces deux phases d'annotation, nous introduisons tout d'abord les mesures utilisées pour évaluer l'accord entre les annotateurs.

#### 3.2.4.1 Mesures d'accord inter-annotateurs pour la reconnaissance d'entités nommées

Pour la tâche de reconnaissance d'entités nommées, les accords inter-annotateurs ont été mesurés avec trois indicateurs : (1) pourcentage d'accord, (2) Kappa de Fleiss (défini dans la section 3.2.3.1) et (3) F-mesure (ou F1-Score). Le pourcentage d'accord est calculé au niveau du document pour chaque étiquette, tandis que le Kappa de Fleiss et le F1-score sont calculés au niveau des mots pour chaque étiquette. Bien que le Kappa de Cohen ou de Fleiss soient souvent considérés comme les mesures standards pour mesurer l'accord inter-annotateurs, Hripcsak and Rothschild (2005) ont montré que le Kappa n'est pas la mesure appropriée pour la tâche de reconnaissance d'entités nommées. En effet, le Kappa nécessite un nombre de cas négatifs pour être calculé, ce qui est inconnu dans le cas des entités nommées. Les entités nommées sont des séquences de mots et le nombre d'éléments à considérer lors de l'annotation d'un texte n'est pas fixe. Le Kappa de Cohen peut être calculé au niveau mot pour un texte donné. Cependant, le nombre de cas négatifs (les mots n'ayant pas d'annotations) sera beaucoup plus élevé que le nombre de cas positifs. La mesure du Kappa sera donc réalisée sur des données très déséquilibrées et nous retomberons dans le paradoxe de Kappa, comme introduit dans la section 3.2.3.1. C'est pour cette raison que la F-mesure, qui ne nécessite pas le nombre de cas négatifs, est généralement reconnue comme une meilleure façon de mesurer l'accord entre annotateurs pour les tâches d'annotation d'entités nommées. Nous mesurons tout de même le Kappa de Fleiss dans un but de comparaison avec la F-mesure.

**Pourcentage d'accord** Pour cette tâche, le pourcentage d'accord correspond au nombre de CRH où tous les annotateurs sont d'accord sur la présence (ou l'absence) d'une étiquette dans le CRH, peu importe le nombre de fois où apparaît l'étiquette et sans prendre en compte le segment de texte annoté pour ces annotations. Cette mesure est donc réalisée au niveau document et nous formulons ce pourcentage d'accord comme suit :

$$P_{oj} = \frac{n_j}{N}$$

où $P_{oj}$ est le pourcentage d'accord pour l'étiquette $j$, $n_j$ est le nombre de CRH où l'étiquette $j$ apparaît au moins une fois et $N$ est le nombre total de CRH annotés.

**F1-score (ou F-mesure)** est la moyenne harmonique entre la précision et le rappel. Pour chaque étiquette, cette métrique est calculée pour chaque paire d'annotateurs à partir de l'ensemble des mots d'un document.

$$F = \frac{2 \times \text{précision} \times \text{rappel}}{\text{précision} + \text{rappel}}$$

$$\text{précision} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

$$\text{rappel} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

où VP est le nombre de vrais positifs, FP le nombre de faux positifs et FN le nombre de faux négatifs.

**F1-score - *Exact match* (EM-F1)** est réalisée au niveau de chaque annotation et non à partir de l'ensemble des mots d'un document comme précédemment. Elle permet de vérifier la correspondance exacte de l'annotation entre deux annotateurs, c'est-à-dire que dans un intervalle de texte convenu, les annotateurs ont attribué la même étiquette. Le score final est obtenu en faisant la moyenne des scores entre chaque paire d'annotateurs.

**F1-score - *Partial match boundaries* (PMB-F1)** est également réalisée au niveau de chaque annotation. Cette mesure est plus souple que l'*Exact match* et vérifie la correspondance partielle de l'annotation entre deux annotateurs, c'est-à-dire que le début ou la fin de l'intervalle de texte de l'annotation correspondent entre les annotateurs et qu'ils ont attribué la même étiquette. Le score final est obtenu en faisant la moyenne des scores entre chaque paire d'annotateurs.

**F1-score - *Partial match tokens* (PMT-F1)** est également réalisée au niveau de chaque annotation. Cette mesure est plus souple que le *Partial match boundaries* et vérifie la correspondance partielle de l'annotation entre deux annotateurs, c'est-à-dire qu'au moins un mot dans l'intervalle de texte est en commun entre les annotations réalisées

par les annotateurs et qu'ils ont attribué la même étiquette. Le score final est obtenu en faisant la moyenne des scores entre chaque paire d'annotateurs.

#### 3.2.4.2 Première phase d'annotation : accords inter-annotateurs et adjudication

De manière générale, les mêmes cas de discordances sont observés sur l'ensemble des variables : erreur d'étiquette lors de l'annotation, oubli d'une annotation, annotation isolée d'un annotateur plus attentif que les autres ou quelques erreurs systématiques dû à un manque de précisions dans les consignes du guide d'annotation. Pour détailler les discordances plus spécifiques à certaines variables, nous proposons deux tableaux pour chaque bloc d'annotation. Le premier rassemble le nombre d'annotations réalisées par chaque annotateur pour chaque variable et étiquette. Le second tableau présente les mesures d'accord inter-annotateurs pour chaque variable et étiquette.

Pour le bloc 1, les annotations réalisées et les accords inter-annotateurs sont présentés dans les tableaux 3.5 et 3.6 respectivement. Il y a globalement très peu d'annotations pour la plupart des étiquettes. Pour les facteurs déclenchant de l'ICA, malgré une faible quantité d'annotations par étiquette, l'accord est faible entre les annotateurs avec un score F1 dépassant rarement 0,30. Lorsque les annotateurs sont en accord sur la présence d'une étiquette à l'échelle du document, cet accord n'est pas forcément conservé sur les empans de texte annotés dans les documents. Parmi les points de discordances, la notion d'exacerbation de BPCO est considéré par certains annotateurs comme étant nécessairement un facteur déclenchant infectieux de l'ICA. Un autre exemple de discordance est la notion de douleur angineuse où certains annotateurs l'ont associé à tort à un facteur déclenchant ischémique de l'ICA. Pour le traitement habituel, le pourcentage d'accord à 62 % est notamment dû à des oublis de certains annotateurs. Au niveau des annotations dans les documents, l'accord est bon avec un Kappa de Fleiss à 0,76 et un score F1 à 0,78.

| Bloc 1 - Facteur déclenchant, 1er épisode ICA et traitement | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Variable | Étiquette | A1 | A2 | A3 | A4 | A5 | A6 | A7 |
| Facteur déclenchant de l'ICA | Trouble du rythme | 8 (5) | 13 (7) | 4 (3) | 7 (4) | 6 (6) | 6 (6) | 3 (2) |
| | Ischémique | 3 (2) | 2 (1) | 4 (2) | 2 (2) | 5 (2) | 3 (3) | 5 (3) |
| | Poussée hypertensive | 9 (6) | 4 (3) | 9 (5) | 8 (6) | 9 (5) | 4 (3) | 6 (4) |
| | Infectieux | 19 (16) | 29 (13) | 35 (14) | 33 (14) | 26 (13) | 12 (8) | 27 (16) |
| | Régime ou traitement | 13 (8) | 3 (3) | 5 (5) | 5 (5) | 8 (6) | 4 (4) | 6 (6) |
| | Autre ou inconnu | 19 (15) | 17 (13) | 29 (22) | 29 (21) | 34 (24) | 13 (12) | 31 (21) |
| Premier épisode ICA | Oui | 5 (3) | 4 (2) | 3 (2) | 2 (1) | 1 (1) | 0 (0) | 1 (1) |
| | Non | 12 (10) | 18 (15) | 11 (9) | 8 (8) | 8 (7) | 5 (5) | 10 (10) |
| Traitement habituel | Oui | 65 (39) | 33 (31) | 33 (33) | 30 (30) | 32 (32) | 32 (29) | 35 (35) |
| Arrêt cardiaque à l'admission | Oui | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| Dans les cases : Nombre total d'annotations (Nombre de CR associés) | | | | | | | | |

TABLEAU 3.5 : Annotations réalisées sur 50 CRH pour le bloc 1.

| Bloc 1 - Facteur déclenchant, 1er épisode ICA et traitement | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | Étiquette | % Accord | Fleiss | F1 | EM-F1 | PMB-F1 | PMT-F1 |
| Facteur déclenchant de l'ICA | Trouble du rythme | 42 (84 %) | 0,22 | 0,19 | 0,11 | 0,22 | 0,25 |
| | Ischémique | 46 (92 %) | 0,22 | 0,23 | 0,08 | 0,3 | 0,31 |
| | Poussée hypertensive | 44 (88 %) | 0,29 | 0,34 | 0,13 | 0,4 | 0,43 |
| | Infectieux | 36 (72 %) | 0,27 | 0,32 | 0,14 | 0,38 | 0,45 |
| | Régime ou traitement | 40 (80 %) | 0,24 | 0,37 | 0,16 | 0,34 | 0,44 |
| | Autre ou inconnu | 17 (34 %) | 0,18 | 0,2 | 0,09 | 0,21 | 0,25 |
| Premier épisode ICA | Oui | 45 (90 %) | 0,07 | 0,13 | 0,09 | 0,16 | 0,16 |
| | Non | 34 (68 %) | 0,28 | 0,26 | 0,06 | 0,27 | 0,34 |
| Traitement habituel | Oui | 31 (62 %) | 0,76 | 0,78 | 0,22 | 0,71 | 0,78 |
| Arrêt cardiaque à l'admission | Oui | 50 (100 %) | 0 | 0 | 0 | 0 | 0 |

TABLEAU 3.6 : Accords inter-annotateurs pour le bloc 1.

Les annotations réalisées et les accords inter-annotateurs mesurés pour le bloc 2 sont présentés dans les tableaux 3.7 et 3.8 respectivement. Les cardiopathies causales ischémique et valvulaire sont les plus nombreuses, on les retrouve dans environ 25 % des CRH. Ces deux cardiopathies causales sont également celles où les annotateurs sont le plus en accord avec un score F1 modéré à 0,46. Pour les autres étiquettes, les annotateurs ne sont pas unanimes sur leur annotation dans les documents. En effet, les cardiopathies causales rythmique, autre et non connu sont identifiées par seulement 2 ou 3 annotateurs sur plusieurs CRH. Pour les types d'ICA, l'insuffisance cardiaque gauche décompensée et les œdème aigu pulmonaire (OAP) sont les types les plus courants. Les accords inter-annotateurs sont faibles à modérés avec des scores F1 allant de 0,14 à 0,44. Les causes de ces discordances sont similaires à celles observées dans les autres blocs, à savoir des CRH annotés seulement par un ou deux annotateurs pour certaines étiquettes. Cela transparaît clairement dans les annotations de l'étiquette sur l'insuffisance cardiaque gauche

décompensée, où le pourcentage d'accord est seulement de 8 %.

| Bloc 2 - Cardiopathie causale et type d'ICA | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Variable | Étiquette | A1 | A2 | A3 | A4 | A5 | A6 | A7 |
| Antécédent d'hypertension artérielle (HTA) | Oui | 6 (5) | 24 (20) | 24 (20) | 11 (8) | 7 (5) | 7 (7) | 11 (8) |
| Cardiopathie causale | Ishémique | 67 (17) | 35 (18) | 32 (20) | 30 (17) | 39 (17) | 13 (12) | 29 (16) |
| | Valvulaire | 48 (16) | 34 (15) | 45 (18) | 36 (16) | 25 (15) | 12 (12) | 18 (12) |
| | Rythmique | 6 (4) | 13 (10) | 16 (12) | 12 (8) | 23 (12) | 7 (7) | 15 (13) |
| | Autre | 8 (6) | 6 (5) | 6 (5) | 8 (6) | 4 (3) | 2 (2) | 4 (3) |
| | Non connu | 7 (5) | 2 (2) | 5 (3) | 2 (2) | 2 (2) | 3 (3) | 3 (3) |
| Type d'ICA | Droite isolée | 1 (1) | 1 (1) | 15 (13) | 17 (10) | 29 (20) | 0 (0) | 2 (1) |
| | Insuffisance cardiaque gauche décompensée | 71 (36) | 78 (36) | 95 (40) | 23 (14) | 101 (40) | 0 (0) | 73 (33) |
| | OAP | 35 (17) | 30 (15) | 29 (16) | 31 (18) | 83 (24) | 48 (38) | 24 (17) |
| | Choc cardiogénique | 7 (2) | 4 (1) | 8 (1) | 8 (2) | 11 (4) | 2 (1) | 5 (1) |
| Dans les cases : Nombre total d'annotations (Nombre de CR associés) | | | | | | | | |

Tableau 3.7 : Annotations réalisées sur 50 CRH pour le bloc 2.

| Bloc 2 - Cardiopathie causale et type d'ICA | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | Étiquette | % Accord | Fleiss | F1 | EM-F1 | PMB-F1 | PMT-F1 |
| Antécédent d'hypertension artérielle (HTA) | Oui | 27 (54 %) | 0,24 | 0,37 | 0,34 | 0,38 | 0,38 |
| Cardiopathie causale | Ischémique | 38 (76 %) | 0,47 | 0,46 | 0,32 | 0,59 | 0,61 |
| | Valvulaire | 37 (74 %) | 0,46 | 0,45 | 0,3 | 0,52 | 0,54 |
| | Rythmique | 36 (72 %) | 0,3 | 0,39 | 0,28 | 0,46 | 0,46 |
| | Autre | 40 (80 %) | 0,2 | 0,28 | 0,2 | 0,29 | 0,29 |
| | Non connu | 42 (84 %) | 0,16 | 0,23 | 0,24 | 0,24 | 0,24 |
| Type d'ICA | Droite isolée | 27 (54 %) | 0,11 | 0,14 | 0,16 | 0,18 | 0,18 |
| | Insuffisance cardiaque gauche décompensée | 4 (8 %) | 0,32 | 0,28 | 0,27 | 0,32 | 0,32 |
| | OAP | 20 (40 %) | 0,2 | 0,44 | 0,39 | 0,5 | 0,51 |
| | Choc cardiogénique | 47 (94 %) | 0,18 | 0,29 | 0,32 | 0,37 | 0,37 |

Tableau 3.8 : Accords inter-annotateurs pour le bloc 2.

Les annotations réalisées et les accords inter-annotateurs mesurés pour le bloc 3 sont présentés dans les tableaux 3.9 et 3.10 respectivement. Les variables du bloc 3 apparaissent très peu dans les CRH. Les accords mesurées sont variables selon les étiquettes, allant d'un accord très faible pour l'antécédent d'insuffisance respiratoire chronique à des accords modérés pour les antécédents de BPCO, de SAOS et d'AVC ou AIT. Pour l'antécédent d'insuffisance respiratoire chronique, les discordances sont notamment liés à la notion d'asthme parfois annotée par certains annotateurs. Un accord modéré a été mesuré pour les diabètes de types 1 et 2 avec quelques annotations isolées de certains annotateurs. Pour l'absence de diabète, les discordances sont principalement dues à des annotations de paragraphes entiers pour indiquer que le diabète n'était pas présent.

| Bloc 3 - Antécédents respiratoires, AVC et diabète | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Variable | Étiquette | A1 | A2 | A3 | A4 | A5 | A6 | A7 |
| Antécédent d'insuffisance respiratoire chronique | Oui | 0 (0) | 2 (1) | 8 (5) | 3 (2) | 7 (2) | 0 (0) | 11 (6) |
| Antécédent BPCO | Oui | 16 (6) | 14 (6) | 13 (6) | 14 (6) | 13 (6) | 3 (3) | 6 (5) |
| Antécédent SAOS | Oui | 3 (1) | 2 (1) | 2 (1) | 1 (1) | 2 (1) | 0 (0) | 1 (1) |
| Antécédent d'AVC ou AIT | Oui | 3 (3) | 3 (3) | 4 (3) | 3 (3) | 0 (0) | 2 (2) | 3 (3) |
| Diabète | Pas de diabète | 2 (2) | 0 (0) | 2 (2) | 0 (0) | 9 (9) | 0 (0) | 7 (7) |
| | Diabète type 1 | 4 (3) | 8 (5) | 4 (3) | 4 (3) | 4 (3) | 1 (1) | 4 (3) |
| | Diabète type 2 | 6 (5) | 10 (7) | 4 (4) | 6 (4) | 5 (5) | 3 (3) | 4 (4) |
| | Diabète autre ou indéterminé | 14 (6) | 0 (0) | 17 (8) | 3 (1) | 14 (6) | 1 (1) | 12 (7) |
| Dans les cases : Nombre total d'annotations (Nombre de CR associés) | | | | | | | | |

Tableau 3.9 : Annotations réalisées sur 50 CRH pour le bloc 3.

| Bloc 3 - Antécédents respiratoires, AVC et diabète | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | Étiquette | % Accord | Fleiss | F1 | EM-F1 | PMB-F1 | PMT-F1 |
| Antécédent d'insuffisance respiratoire chronique | Oui | 41 (82 %) | 0,04 | 0,03 | 0,02 | 0,04 | 0,04 |
| Antécédent BPCO | Oui | 46 (92 %) | 0,41 | 0,43 | 0,35 | 0,54 | 0,56 |
| Antécédent SAOS | Oui | 49 (98 %) | 0,57 | 0,49 | 0,11 | 0,41 | 0,54 |
| Antécédent d'AVC ou AIT | Oui | 47 (94 %) | 0,52 | 0,49 | 0,11 | 0,41 | 0,54 |
| Diabète | Pas de diabète | 41 (82 %) | 0,19 | 0,08 | 0,03 | 0,08 | 0,09 |
| | Diabète type 1 | 46 (92 %) | 0,46 | 0,62 | 0,49 | 0,65 | 0,65 |
| | Diabète type 2 | 42 (84 %) | 0,3 | 0,49 | 0,35 | 0,52 | 0,52 |
| | Diabète autre ou indéterminé | 42 (84 %) | 0,26 | 0,16 | 0,13 | 0,19 | 0,19 |

Tableau 3.10 : Accords inter-annotateurs pour le bloc 3.

Les annotations réalisées et les accords inter-annotateurs mesurés pour le bloc 4 sont présentés dans les tableaux 3.11 et 3.12 respectivement. Les informations cliniques à l'admission sont les variables pour lesquelles nous observons les meilleurs accords inter-annotateurs. Cela est facilité par une définition commune des informations à annoter partagée entre les annotateurs. Ces variables sont donc moins sujettes à l'interprétation par rapport aux variables des autres blocs d'annotation. Les accords sont un peu plus faibles pour le poids et l'IMC car plusieurs poids peuvent être mesurés au cours du séjour (poids à l'entrée, poids sec, poids de sortie, poids après dialyse, etc.) et rapportés dans un même CRH. Pour le tabagisme, très peu d'informations ont été annotées et les accords sont modérés pour les modalités sevré et actif.

| Bloc 4 - Infos cliniques à l'admission et tabagisme | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Variable | Étiquette | A1 | A2 | A3 | A4 | A5 | A6 | A7 |
| Fréquence cardiaque à l'admission | FC | 22 (22) | 21 (21) | 28 (22) | 21 (21) | 28 (23) | 24 (22) | 24 (22) |
| Pression artérielle à l'admission | PA | 24 (24) | 24 (24) | 25 (23) | 25 (25) | 29 (24) | 26 (24) | 26 (25) |
| Poids | Poids | 17 (15) | 11 (11) | 28 (19) | 12 (12) | 29 (18) | 14 (12) | 17 (16) |
| Taille | Taille | 7 (7) | 6 (6) | 7 (7) | 6 (6) | 7 (7) | 6 (6) | 7 (6) |
| IMC | IMC | 2 (2) | 0 (0) | 1 (1) | 0 (0) | 2 (2) | 0 (0) | 1 (1) |
| Tabagisme | Jamais | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 3 (3) | 0 (0) | 0 (0) |
| | Sevré | 7 (6) | 6 (6) | 8 (7) | 6 (6) | 7 (6) | 2 (2) | 6 (6) |
| | Actif | 1 (1) | 1 (1) | 1 (1) | 1 (1) | 3 (2) | 2 (2) | 2 (2) |
| Dans les cases : Nombre total d'annotations (Nombre de CR associés) | | | | | | | | |

Tableau 3.11 : Annotations réalisées sur 50 CRH pour le bloc 4.

| Bloc 4 - Infos cliniques à l'admission et tabagisme | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | Étiquette | % Accord | Fleiss | F1 | EM-F1 | PMB-F1 | PMT-F1 |
| Fréquence cardiaque à l'admission | FC | 45 (90 %) | 0,65 | 0,62 | 0,33 | 0,77 | 0,86 |
| Pression artérielle à l'admission | PA | 47 (94 %) | 0,68 | 0,66 | 0,32 | 0,77 | 0,92 |
| Poids | Poids | 38 (76 %) | 0,39 | 0,4 | 0,14 | 0,44 | 0,59 |
| Taille | Taille | 48 (96 %) | 0,68 | 0,64 | 0,3 | 0,71 | 0,8 |
| IMC | IMC | 48 (96 %) | 0,28 | 0,14 | 0,1 | 0,19 | 0,19 |
| Tabagisme | Jamais | 47 (94 %) | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| | Actif | 45 (90 %) | 0,58 | 0,62 | 0,38 | 0,73 | 0,73 |
| | Sevré | 48 (96 %) | 0,27 | 0,44 | 0,17 | 0,62 | 0,63 |

Tableau 3.12 : Accords inter-annotateurs pour le bloc 4.

Les annotations réalisées et les accords inter-annotateurs mesurés pour le bloc 5 sont présentés dans les tableaux 3.13 et 3.14 respectivement. Les différents antécédents apparaissent peu dans les CRH et les accords inter-annotateurs sont faibles. On retrouve globalement les mêmes schémas de discordances que dans les autres blocs avec des annotations isolées par certains annotateurs ou des annotations partagées par seulement un sous ensemble des annotateurs et ne font pas l'unanimité. Plus précisément, pour l'antécédent de dépression, les discordances portent sur l'annotation des antidépresseurs comme Mianserine ou Norset par un seul des annotateurs. Pour l'antécédent de cancer, certains cancers comme la leucémie lymphoïde chronique (LLC) ont été identifiés par un seul des annotateurs.

| Bloc 5 - Antécédent ACFA, 1ère FEVG et autres comorbidités | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Variable | Étiquette | A1 | A2 | A3 | A4 | A5 | A6 | A7 |
| Antécédent trouble du rythme | Oui | 16 (10) | 10 (9) | 12 (9) | 11 (11) | 20 (10) | 9 (9) | 9 (9) |
| Antécédent dépression | Oui | 0 (0) | 2 (2) | 2 (2) | 1 (1) | 8 (6) | 1 (1) | 2 (2) |
| Antécédent de troubles cognitifs | Oui | 14 (6) | 9 (6) | 10 (7) | 8 (6) | 10 (4) | 1 (1) | 7 (6) |
| Antécédent de cancer | Oui | 27 (9) | 8 (6) | 8 (6) | 9 (6) | 16 (7) | 1 (1) | 7 (7) |
| AC/FA à l'admission | Oui | 4 (3) | 9 (9) | 11 (9) | 11 (9) | 10 (7) | 0 (0) | 7 (7) |
| FEVG | $\leq 40$ | 34 (12) | 9 (8) | 15 (8) | 10 (7) | 23 (8) | 11 (7) | 8 (8) |
| | $41 \leq$ FEVG $\leq 49$ | 3 (3) | 3 (3) | 4 (4) | 4 (4) | 4 (4) | 3 (3) | 4 (4) |
| | $\geq 50$ | 11 (8) | 8 (8) | 11 (10) | 9 (9) | 14 (10) | 6 (6) | 11 (11) |
| Dans les cases : Nombre total d'annotations (Nombre de CR associés) | | | | | | | | |

Tableau 3.13 : Annotations réalisées sur 50 CRH pour le bloc 5.

| Bloc 5 - Antécédent ACFA, 1ère FEVG et autres comorbidités | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | Étiquette | % Accord | Fleiss | F1 | EM-F1 | PMB-F1 | PMT-F1 |
| Antécédent trouble du rythme | Oui | 42 (84 %) | 0,33 | 0,46 | 0,36 | 0,58 | 0,59 |
| Antécédent dépression | Oui | 44 (88 %) | 0,16 | 0,3 | 0,19 | 0,36 | 0,36 |
| Antécédent de troubles cognitifs | Oui | 43 (86 %) | 0,36 | 0,39 | 0,25 | 0,44 | 0,49 |
| Antécédent de cancer | Oui | 42 (86 %) | 0,33 | 0,34 | 0,21 | 0,45 | 0,46 |
| AC/FA à l'admission | Oui | 37 (84 %) | 0,2 | 0,16 | 0,09 | 0,21 | 0,22 |
| FEVG | $\leq 40$ | 43 (86 %) | 0,29 | 0,43 | 0,17 | 0,38 | 0,52 |
| | $41 \leq$ FEVG $\leq 49$ | 47 (94 %) | 0,73 | 0,63 | 0,3 | 0,56 | 0,79 |
| | $\geq 50$ | 43 (86 %) | 0,64 | 0,58 | 0,29 | 0,51 | 0,67 |

Tableau 3.14 : Accords inter-annotateurs pour le bloc 5.

Les accords inter-annotateurs obtenus sur les 50 CRH supplémentaires annotés par PCDB, SH et MW sont globalement similaires. On peut tout de même noter quelques hausses d'accord sur certaines variables comme l'antécédent d'insuffisance respiratoire chronique avec une F-mesure de 0,33 (+0,30), l'IMC avec une F-mesure de 0,69 (+0,55) et le tabagisme sevré avec une F-mesure de 0,65 (+0,44). Les hausses d'accord observées montrent que ces trois annotateurs abordent ces variables de manière très différente par rapport aux quatre autres annotateurs.

#### 3.2.4.3 Deuxième phase d'annotation : statistiques du corpus

Les tableaux 3.15 à 3.19 rassemblent, pour chaque bloc d'annotation, les annotations réalisées sur les 500 CRH pour chaque étiquette. Dans le bloc 1 présenté dans le tableau 3.15, le traitement habituel est la variable la plus représentée avec 330 CRH annotés avec cette information. Les autres variables sont plus nuancées. L'information portant sur le premier épisode d'ICA connu du patient est annoté dans 25 % des CRH, tandis que les facteurs déclenchant de l'ICA sont en moyenne moins présents. Quant à la notion d'arrêt cardiaque à l'admission du patient, cette variable n'est pas du tout renseignée dans les CRH de la population de l'étude.

| Bloc 1 - Facteur déclenchant, 1er épisode ICA et traitement | | | |
|---|---|---|---|
| Variable | Étiquette | # annotations | # CRH |
| Facteur déclenchant de l'ICA | Trouble du rythme | 137 | 83 |
| | Ischémique | 44 | 28 |
| | Poussée hypertensive | 72 | 45 |
| | Infectieux | 292 | 144 |
| | Valve | 33 | 24 |
| | Régime ou traitement | 69 | 44 |
| | Autre ou inconnu | 248 | 145 |
| Premier épisode ICA | Oui | 34 | 25 |
| | Non | 156 | 108 |
| Traitement habituel | Oui | 379 | 330 |
| Arrêt cardiaque à l'admission | Oui | 4 | 3 |

Tableau 3.15 : Annotations finales des 500 CRH pour le bloc 1.

Pour les annotations du bloc 2 présenté dans le tableau 3.16, les variables sont assez fréquentes. La notion d'antécédent d'hypertension artérielle et le type d'ICA gauche décompensée sont retrouvés dans la moitié des CRH annotés. Un peu moins fréquents, les

| Bloc 2 - Cardiopathie causale et type d'ICA | | | |
|---|---|---|---|
| Variable | Étiquette | # annotations | # CRH |
| Antécédent d'hypertension artérielle (HTA) | Oui | 341 | 248 |
| Cardiopathie causale | Ischémique | 414 | 152 |
| | Valvulaire | 434 | 131 |
| | Rythmique | 275 | 166 |
| | Autre | 121 | 71 |
| | Non connu | 35 | 17 |
| Type d'ICA | Droite isolée | 236 | 107 |
| | Insuffisance cardiaque gauche décompensée | 976 | 316 |
| | OAP | 480 | 170 |
| | Choc cardiogénique | 73 | 26 |

Tableau 3.16 : Annotations finales des 500 CRH pour le bloc 2.

cardiopathies causales ischémique, valvulaire et rythmique sont annotées dans environ un tiers des CRH.

| Bloc 3 - Antécédents respiratoires, AVC et diabète | | | |
|---|---|---|---|
| Variable | Étiquette | # annotations | # CRH |
| Antécédent d'insuffisance respiratoire chronique | Oui | 68 | 42 |
| Antécédent BPCO | Oui | 92 | 63 |
| Antécédent SAOS | Oui | 36 | 32 |
| Antécédent d'AVC ou AIT | Oui | 81 | 67 |
| Diabète | Pas de diabète | 6 | 6 |
| | Diabète type 1 | 8 | 5 |
| | Diabète type 2 | 111 | 100 |
| | Diabète autre ou indéterminé | 97 | 55 |

Tableau 3.17 : Annotations finales des 500 CRH pour le bloc 3.

Les antécédents respiratoires et d'AVC du bloc 3 dans le tableau 3.17 sont annotés dans seulement 10 % des CRH. Pour ce qui est du diabète, seul le diabète de type 2 est souvent annoté.

| Bloc 4 - Infos cliniques à l'admission et tabagisme | | | |
|---|---|---|---|
| Variable | Étiquette | # annotations | # CRH |
| Fréquence cardiaque à l'admission | FC | 334 | 277 |
| Pression artérielle à l'admission | PA | 292 | 267 |
| Poids | Poids | 193 | 158 |
| Taille | Taille | 77 | 73 |
| IMC | IMC | 42 | 40 |
| Tabagisme | Jamais | 13 | 13 |
| | Actif | 29 | 28 |
| | Sevré | 71 | 69 |

Tableau 3.18 : Annotations finales des 500 CRH pour le bloc 4.

Dans le tableau 3.18, les informations cliniques à l'admission, que l'on peut aussi retrouver dans les données structurées, sont des variables très souvent annotés. La fréquence cardiaque et la pression artérielle sont annotées dans plus de 50 % des CRH, tandis que le poids est annoté dans environ 30 % des cas. Le tabagisme est globalement très peu présent dans les CRH avec un total de 110 annotations sur les 500 CRH.

| Bloc 5 - Antécédent ACFA, 1ère FEVG et autres comorbidités | | | |
|---|---|---|---|
| Variable | Étiquette | # annotations | # CRH |
| Antécédent troubles du rythme | Oui | 206 | 156 |
| Antécédent dépression | Oui | 60 | 41 |
| Antécédent troubles cognitifs | Oui | 142 | 69 |
| Antécédent de cancer | Oui | 142 | 88 |
| AC/FA à l'admission | Oui | 215 | 143 |
| FEVG | $\leq 40$ | 136 | 80 |
| | $41 \leq$ FEVG $\leq 49$ | 23 | 19 |
| | $\geq 50$ | 148 | 115 |

Tableau 3.19 : Annotations finales des 500 CRH pour le bloc 5.

Enfin, pour le bloc 5 présenté dans le tableau 3.19, les variables d'antécédent de troubles du rythme et d'AC/FA à l'admission sont présentes dans environ 35 % des CRH. Les antécédents de dépression, de cancer et de troubles cognitifs sont moins représentés avec en moyenne entre 9 à 15 % des CRH annotés avec ces informations. La FEVG est globalement retrouvé dans environ 50 % des documents.

De manière générale, les étiquettes annotées dans les CRH sont plutôt des évènements rares. En effet, pour la plupart, lorsqu'une étiquette est présente dans un CRH, il y a rarement plus d'une annotation pour cette étiquette. Pour les variables ayant plusieurs

étiquettes contradictoires dans un même document, un ensemble de règles a été défini afin de prendre une décision sur l'étiquette finale qui sera associée, pour cette variable, à ce document.

## 3.3 Déterminants sociaux de santé

Les conditions sociales et économiques interagissent et influencent l'état de santé des populations. Malgré une augmentation continue de l'espérance de vie de manière générale, les inégalités persistent et s'accroissent entre les catégories sociales tout au long de la vie (Chetty et al., 2016). Les progrès de la médecine peuvent prolonger la vie des personnes atteintes de maladies graves, mais la gestion thérapeutique ne doit pas être déconnectée de l'environnement socio-économique dans lequel vivent les patients (Marmot, 2005). En effet, les trajectoires des maladies chroniques sont influencées par les comportements et les expositions. Leur combinaison résulte d'un processus socio-biologique complexe qui détermine l'impact individuel et sociétal des conditions de santé (Magnan, 2017).

Les déterminants sociaux de la santé (DSS) sont les conditions dans lesquelles les personnes grandissent, vivent et travaillent. Ils influencent leur qualité de vie et leur santé (Marmot, 2005). Les informations sur les DSS sont en partie recueillies lors des soins cliniques et sont disponibles dans le dossier patient informatisé à la fois sous forme de données structurées et de textes cliniques non structurées. Cependant, les textes cliniques fournissent des détails plus complexes et subtils sur la plupart des DSS que les données structurées. Pour une utilisation secondaire à grande échelle, telle que des études cliniques, le recueil manuel d'informations sur les DSS à partir des textes cliniques est impossible et nécessite une extraction automatique à l'aide d'outils de TAL. Les méthodes d'extraction d'informations mettent en œuvre des méthodes à base d'apprentissage supervisé nécessitant des corpus annotés.

Ces dernières années, l'extraction automatique des DSS a fait l'objet de nombreuses études dans la langue anglaise, comme le présente la revue systématique de Patra et al. (2021). Les DSS qui apparaissent le plus sont le statut tabagique, les problèmes liés à la consommation de substances, à l'alcool et au logement. En effet, la consommation d'alcool, le tabagisme et la toxicomanie sont les principaux facteurs de risque modifiables des maladies chroniques les plus répandues, tandis que l'instabilité du logement peut avoir des effets négatifs sur la santé et le bien-être et limiter l'accès aux soins.

De nombreux corpus cliniques annotés avec des DSS ont été proposés pour la langue anglaise. L'extraction des DSS à partir des textes cliniques peut être abordée via plusieurs tâches. Certains corpus sont annotés au niveau document et l'extraction se fait par le biais d'une tâche de classification de texte. Parmi ces corpus, certains ciblent des DSS spécifiques, tels que la consommation de substances (Gehrmann et al., 2018; Feller et al.,

2018c), l'obésité (Gehrmann et al., 2018), le logement (Feller et al., 2018c; Chapman et al., 2021) ou les facteurs liés à la santé sexuelle (Feller et al., 2018c) tandis que d'autres ont une portée beaucoup plus large, comprenant la race, le statut matrimonial, la profession, les revenus financiers et bien d'autres (Yu et al., 2021a, 2022; Han et al., 2022).

D'autres corpus abordent l'annotation des DSS à travers des empans de texte incluant des relations pour extraire manière plus précise les informations. Wang et al. (2015b) utilise un schéma basé sur les relations pour extraire les problèmes liés à la consommation de substances. Yetisgen et al. (2016) et Yetisgen and Vanderwende (2017) annotent avec un schéma basé sur les événements pour extraire 13 DSS (utilisation de substances, situation de vie, etc.). Reeves et al. (2021) utilise des empans de texte et des attributs pour extraire huit DSS, où l'attribut est une assertion (présent, absent ou incertain). D'autres corpus de textes cliniques ont été proposés dans le cadre de campagnes d'évaluation avec des corpus partagés pour étudier l'extraction des DSS par le biais du TAL tels que le challenge i2b2 2006 sur l'identification du statut tabagique des patients, avec un corpus de 502 textes cliniques (Uzuner et al., 2008b), et plus récemment, le challenge n2c2 2022 explorant l'extraction des DSS à partir de 4 405 sections d'antécédents sociaux issus de textes cliniques, comprenant des informations sur l'alcool, la drogue, le tabac, l'emploi et la situation de vie (Lybarger et al., 2023).

Dans les langues autres que l'anglais, peu de ressources sont disponibles. L'extraction du statut tabagique à partir de textes narratifs cliniques a été étudiée en espagnol (Figueroa et al., 2014), en finnois (Karlsson et al., 2021), en suédois (Caccamisi et al., 2020) et dans un contexte bilingue coréen-anglais (Bae et al., 2021). Aucun des corpus utilisés dans ces travaux n'a été partagé avec la communauté. À ce jour, l'extraction des DSS à partir de textes cliniques en français n'a pas été étudiée.

Pour combler ce manque de ressources pour le français et étudier l'extraction automatique des DSS dans les comptes rendus cliniques, nous avons construit un corpus composé de paragraphes issus de 1 700 textes cliniques de l'entrepôt de données du CHU de Nantes. Ces paragraphes sont annotés avec un schéma basé sur des entités et des relations pour 13 DSS : conditions de vie, statut matrimonial, descendance, statut d'emploi, profession, consommation de tabac, consommation d'alcool, consommation de substances, logement, activité physique, revenu, niveau d'éducation et origine ethnique. Dans un premier temps, nous décrivons les étapes de préparation en amont de l'annotation du corpus : la sélection des données et leur prétraitement, le choix de l'outil d'annotation et la définition du schéma d'annotation. Nous présentons ensuite le processus d'annotation du corpus

comportant les phases successives d'annotation et d'adjudication. Enfin, nous donnons quelques statistiques sur le corpus annoté.

### 3.3.1 Préparation de l'annotation

La mise en place de ce corpus suit trois étapes identiques à celles du corpus élaboré pour le projet GAVROCHE. Nous décrivons tout d'abord la constitution du corpus à partir de l'entrepôt de données biomédicales du CHU de Nantes ainsi que les prétraitements effectués sur les comptes rendus. Nous présentons ensuite l'outil d'annotation et le schéma d'annotation avec une description des entités et relations qui le compose.

#### 3.3.1.1 Constitution du corpus et prétraitements

Le corpus a été construit à partir de comptes rendus cliniques de l'entrepôt de données biomédicales du CHU de Nantes. Un total de 1 144 443 comptes rendus hospitaliers ont été sélectionnés et extraits de l'entrepôt selon les critères d'inclusion suivants :

- Femme ou homme d'âge $\geq$ 18 ans,
- CRH provenant de séjours entre le 01/01/2018 et le 01/06/2022.

Le seul critère de non-inclusion est les CRH de patients ayant manifestés leur volonté de ne pas voir leurs données utilisées à des fins de recherche par l'établissement.

Les comptes-rendus hospitaliers sont catégorisés par service et spécialité. Ils peuvent avoir différents formats : certains sont totalement non structurés (comme certaines lettres de liaison, par exemple), d'autres sont semi-structurés car organisés en sections, telles que « Antécédents », « Traitement habituel », « Mode de vie », etc. Parmi les 1 144 443 comptes rendus extraits de l'entrepôt de données, nous avons concentré nos efforts sur les comptes rendus semi-structurés et plus particulièrement sur deux types de comptes rendus : les comptes rendus de consultation et les comptes rendus d'hospitalisation. Ces deux types de comptes rendus couvrent de nombreuses spécialités médicales, ce qui nous permet de disposer de 206 973 comptes rendus comprenant différents profils de patients. Un dernier critère de sélection a été appliqué pour garder uniquement les comptes rendus contenant explicitement une section « Mode de vie », pour un total de 32 666 comptes-rendus. Afin d'éviter d'avoir à lire un document entier lors de l'annotation, les sections « Mode de vie » ont été extraites des comptes-rendus à l'aide d'un système à base de règles, en analysant les titres de sections dans les documents. Finalement, 1 700 sections

« Mode de vie » ont été sélectionnées et font l'objet de cette annotation. Toutes ces étapes sont synthétisées dans la figure 3.2.

#### 3.3.1.2 Définition du schéma d'annotation

Nous avons tout d'abord défini la liste de DSS à partir des déterminants les plus présents dans la littérature scientifique. Cette liste a ensuite été affinée avec l'aide du Dr Cyrille Delpierre dont le domaine de recherche principal porte sur les inégalités sociales dans la santé. La liste de DSS retenus pour l'annotation est présentée dans le tableau 3.21.

Comme nous l'avons énoncé en introduction de ce sous-chapitre, les DSS peuvent être annotés de différentes façons à travers différentes tâches. Nous avons fait le choix d'une annotation basée sur un schéma composé d'entités et de relations pour deux raisons. La première est la fine granularité des informations annotées. En effet, l'annotation en entités et relations permet de récupérer des informations précises, telles que des quantités pour les consommations de substances ou encore la temporalité des DSS annotés. Par exemple, pour le tabagisme, au-delà de la catégorisation (actif — sevré — aucun) que l'on retrouve habituellement pour ce DSS, nous pouvons également récupérer la quantité consommée ainsi que la durée d'exposition. La seconde raison de ce choix de schéma est la modulabilité du corpus. Un corpus annoté en entités et relations peut facilement être converti pour un usage à travers différentes tâches, telles que la classification multi-étiquettes ou le format instructions pour les modèles génératifs. Un exemple d'annotation en entités et relations pour la consommation d'alcool est illustré dans la figure 3.3.

À partir de la liste de DSS, nous avons établi une liste d'entités et de relations à annoter dans les sections « Mode de vie ». Le schéma d'annotation est composé de 25 entités en lien direct avec les DSS qui sont présentées dans le tableau 3.21. Les relations et les entités propres aux relations sont définies dans le tableau 3.20. Pour chaque relation, une liste de paires d'entités `(entité source - entité cible)` est définie. Certaines relations sont obligatoirement présentes lorsque certaines entités sont annotées. Par exemple, les entités `Tabagisme`, `Alcool` et `Drogue` sont systématiquement reliées à une entité `StatusTime` avec la relation `Status`. De même pour l'entité `Descendance` qui est toujours relié à une entité `Type` afin de spécifier si la descendance du patient correspond à ses enfants ou ses petits-enfants.

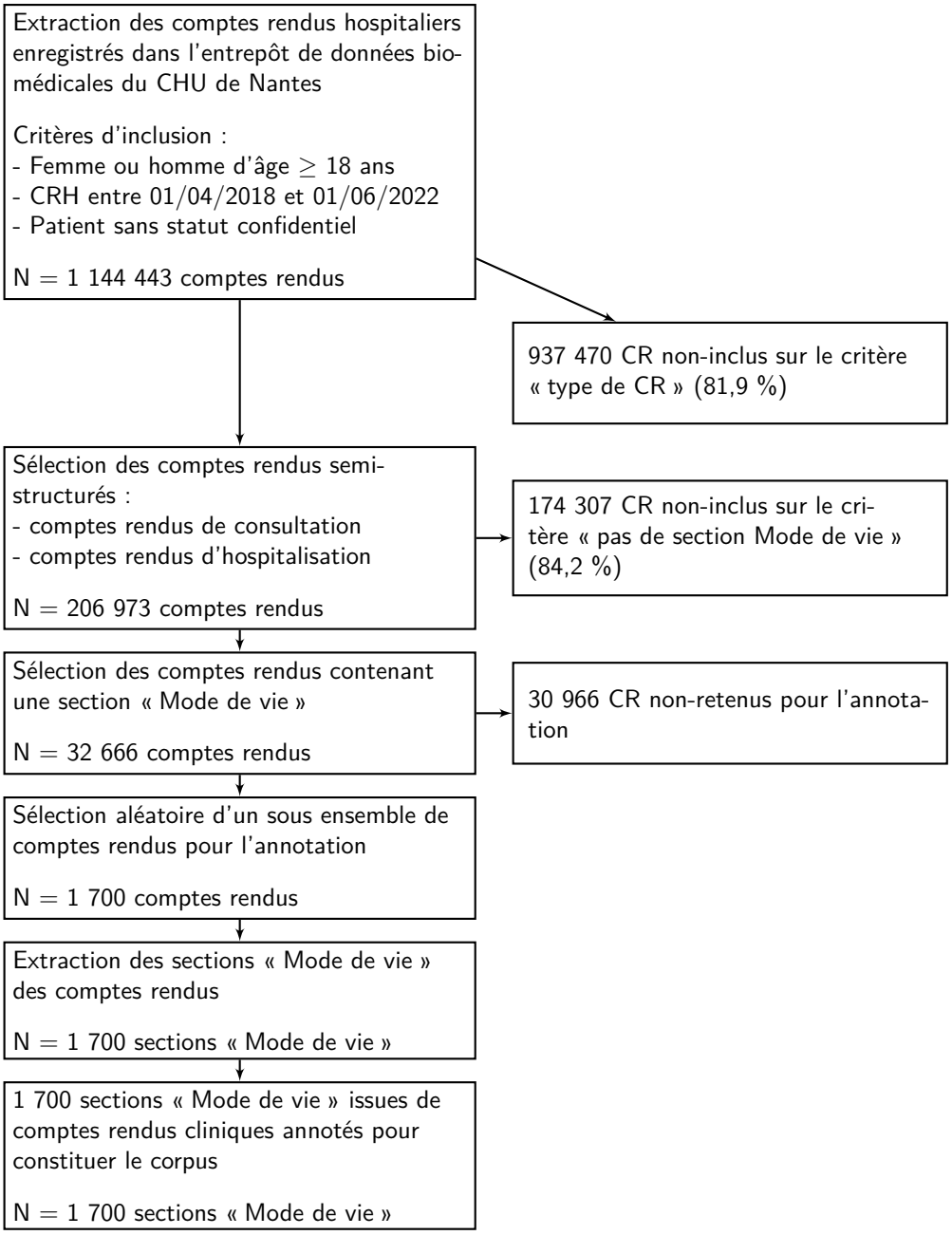


FIGURE 3.2 : Flow-chart de la sélection des comptes rendus hospitaliers dans l'entrepôt de données biomédicales du CHU de Nantes.

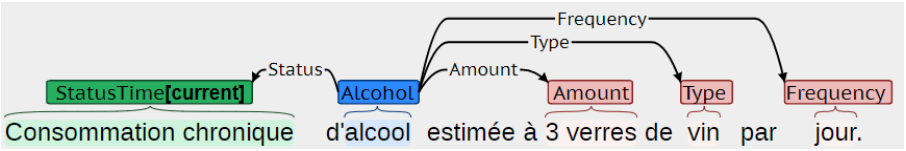


FIGURE 3.3 : Exemple d'annotation pour la consommation d'alcool.

| **Relation** | **Entités associés et attributs** | **Définitions** | **Entités impliquées** |
|---|---|---|---|
| Status | StatusTime<br>- Actif<br>- Sevré<br>- Jamais | Définit le statut d'une consommation de tabac, d'alcool ou de drogues | **Tabagisme - StatusTime**<br>**Alcool - StatusTime**<br>**Drogue - StatusTime** |
| Quantité | Quantité | Associe une quantité à une consommation ou permet de préciser le nombre d'enfants dans la descendance<br>Unités de mesure : nb. verres ; nb. cigarettes ; nb. d'enfants ; grammes | Descendance - Quantité<br>Tabagisme - Quantité<br>Alcool - Quantité<br>Drogue - Quantité |
| Durée | Durée | Durée d'exposition. | Domicile :Non - Durée<br>Tabagisme - Durée<br>Alcool - Durée<br>Drogue - Durée |
| Fréquence | Fréquence | Donne la fréquence d'un DSS, par exemple l'activité physique. Cette relation est notamment utilisée lorsqu'il n'y a pas de quantité précise indiqué avec la relation Quantité. | Tabagisme - Fréquence<br>Alcool - Fréquence<br>Drogue - Fréquence<br>Activité physique - Fréquence |
| Antécédent | Antécédent | Permet de retrouver la date d'apparition d'un évènement | * - Antécédent |
| Type | Type | Permet de préciser certaines entités un large comme le type de descendance, le type d'alcool consommé | **Descendance - Type**<br>Tabagisme - Type<br>Alcool - Type<br>Drogue - Type |

TABLEAU 3.20 : Relations et entités propres aux relations du corpus de déterminants sociaux de santé. Les entités impliquées en gras indiquent que lorsque l'entité source est annotée, il est obligatoire d'avoir une relation avec l'entité cible. Par exemple, lorsque la consommation d'alcool est annotée, le statut de cette consommation (active, sevré ou aucune) doit être systématiquement annoté comme illustré dans la figure 3.3. * signifie que n'importe quelle entité peut être reliée.

### 3.3.2 Annotation du corpus

Contrairement à ce que nous avons décrit sur l'annotation des CRH du projet GAVROCHE, l'annotation du corpus sur les déterminants sociaux de santé est plus simple, car ne requière pas de connaissances médicales spécifiques et peut donc être réalisée par des annotateurs non-experts. L'annotation du corpus a été réalisé par trois annotateurs : Dr Pacôme Constant dit Beaufils, Matilde Karakachoff et moi-même. Nous avons utilisé l'outil brat (Stenetorp et al., 2012) pour faire réaliser les annotations. BRAT est moins ergonomique que prodigy mais permettant des annotations plus complexes grâce à un schéma d'annotation plus souple.

Le processus d'annotation s'est déroulé en trois phases avec chaque phase le calcul des

| **Variable (DSS)** | **Entité** | **Description et exemples** |
|---|---|---|
| Conditions de vie | Seul | Est-ce que le patient vit seul ?<br>Exemples : « vit seul » |
| | Cohabitation | Est-ce que le patient vit avec d'autres personnes ?<br>Exemples : « vit avec son épouse » ; « vit en colocation » ; « vit en EHPAD » |
| Statut matrimonial | Célibataire | Le patient est-il décrit comme célibataire ?<br>Exemples : « célibataire » |
| | Marié/en couple | Le patient est-il décrit comme étant marié, pacsé ou avec un partenaire régulier ?<br>Exemples : « vit en couple » |
| | Séparé/divorcé | Le patient est-il décrit comme divorcé, séparé ou a quitté son ancien partenaire ?<br>Exemples : « divorcé » ; « plus de contact avec le père de ses enfants » |
| | Veuf | Le patient est-il décrit comme veuf ?<br>Exemples : « veuf » |
| Descendance | Oui | Le patient est-il décrit comme ayant une descendance ? (enfants, petits-enfants)<br>Exemples : « Deux filles » ; « Mère au foyer » |
| | Non | Le patient est-il décrit comme n'ayant pas de descendance ?<br>Exemples : « pas d'enfant » ; « Enceinte de son premier enfant » |
| Statut d'emploi | Actif | Le patient est-il décrit comme travaillant actuellement ou ayant un statut d'emploi défini comme occupé ?<br>Exemples : « agent de ménage » ; « travaille toujours » ; « en reconversion professionnelle » |
| | Retraité | Est-ce que le patient est retraité ?<br>Exemples : « retraité » ; « ancien magasinier » |
| | Etudiant | Est-ce que le patient est étudiant ?<br>Exemples : « patient étudiant » ; « actuellement lycéen » ; « en première année de BTS » |
| | Sans emploi/chômage | La patient est-il décrit comme sans emploi actuellement ou au chômage ?<br>Exemples : « elle vient de terminer son CDD » ; « il est au chômage » |
| | Autre | Le patient est-il décrit comme étant dans un situation de travail sans indication sur la temporalité, d'invalidité, de congé maladie de longue durée ou de situation irrégulière ?<br>Exemples : « a été courtier dans le textile » ; « ne travaille plus » ; « en invalidité » |
| Profession | - | Quelles sont les professions que le patient a exercé (ou exerce toujours) ?<br>Exemples : « coiffeuse » ; « soudeur » ; « travaille dans le bâtiment » |
| Dernière profession exercée | - | Quelle est la dernière profession exercée par le patient ?<br>Exemples : « plombier, chauffagiste puis carreleur ». On annotera « carreleur » ici. |
| Tabagisme | - | Est-ce que le patient fume ? a fumé ? n'a jamais fumé ?<br>Cette variable est annotée comme un évènement. On annotera ici le terme déclenchant de l'évènement. Le statut du tabagisme (actif, sevré, jamais) est annoté à travers une relation avec l'entité StatusTime.<br>Exemples : « tabagisme » ; « tabac » ; « consommation tabagique » ; « intoxication alcoolotabagique » |
| Consommation d'alcool | - | Est-ce que le patient boit de l'alcool ? a bu ? n'a jamais bu ?<br>Cette variable est annotée comme un évènement. On annotera ici le terme déclenchant de l'évènement. Le statut de la consommation d'alcool est annoté à travers une relation avec l'entité StatusTime.<br>Exemples : « consommation d'alcool » ; « CAD » ; « OH » ; « ethylisme » ; « exogénose » |
| Consommation de substances | - | Est-ce que le patient consomme des substances illicites ? a consommé ? n'a jamais consommé ?<br>Cette variable est annotée comme un évènement. On annotera ici le terme déclenchant de l'évènement. Le statut de la consommation de substances est annoté à travers une relation avec l'entité StatusTime.<br>Exemples : « drogue » ; « cannabis » ; « héroïne » |
| Domicile | Oui | Le patient est-il décrit comme ayant un domicile fixe ?<br>Exemples : « vit dans une maison » ; « appartement » ; « habite » ; « aide à domicile » |
| | Non | Le patient est-il décrit comme n'ayant pas de domicile fixe ?<br>Exemples : « sans domicile fixe » ; « hébergé chez des tiers » |
| Activité physique | Oui | Le patient est-il décrit comme ayant une activité physique ?<br>Exemples : « Fait du vélo » ; « marche 30min quotidiennement » |
| | Non | Le patient est-il décrit comme n'ayant pas d'activité physique ?<br>Exemples : « pas d'activité sportive » ; « performans status 3 » |
| Revenu | - | Quels sont les revenus financiers du patient ? Perçoit-il des prestations sociales ?<br>Exemples : « RSA » ; « AAH » |
| Niveau d'éducation | - | Quels sont les diplômes du patient ?<br>Exemples : « certificat d'étude » ; « BTS informatique » |
| Origine ethnique | - | Quel est le pays de naissance du patient ?<br>Exemples : « originaire d'Érythrée » ; « originaire du Maroc » |

Tableau 3.21 : Listes des déterminants sociaux de santé (DSS) et leurs entités associées.

accords inter-annotateurs : (1) une phase d'annotation préliminaire sur 100 documents pour éprouver le schéma et les règles d'annotation en calculant les premiers accords inter-

| | **Phase 1 3 annotateurs** | | **Phase 2 2 annotateurs** | |
|---|---|---|---|---|
| | *Annotations F-Mesure* | *Mot F-mesure* | *Annotations F-Mesure* | *Mot F-mesure* |
| **Entités** | | | | |
| Conditions de vie : Seul | 0,337 | 0,598 | 0,688 | 0,755 |
| Conditions de vie : Cohabitation | 0,349 | 0,638 | 0,776 | 0,901 |
| Descendance : Oui | 0,334 | 0,746 | 0,550 | 0,654 |
| Descendance : Non | 0,333 | 0,810 | 0,930 | 0,962 |
| Statut matrimonial : Célibataire | 0,730 | 0,627 | 0,880 | 0,839 |
| Statut matrimonial : Marié | 0,468 | 0,747 | 0,341 | 0,727 |
| Statut matrimonial : Divorcé | 0,926 | 0,926 | 0,640 | 0,476 |
| Statut matrimonial : Veuf | 0,882 | 0,795 | 1,000 | 1,000 |
| Statut d'emploi : Actif | 0,443 | 0,334 | 0,667 | 0,651 |
| Statut d'emploi : Retraité | 0,527 | 0,460 | 0,500 | 0,571 |
| Statut d'emploi : Etudiant | 0,400 | 0,509 | 0,824 | 0,640 |
| Statut d'emploi : Sans emploi | 0,531 | 0,776 | 0,545 | 0,863 |
| Statut d'emploi : Autre | 0,207 | 0,520 | 0,143 | 0,762 |
| Profession | 0,726 | 0,869 | 0,602 | 0,781 |
| Dernière profession exercée | 0,541 | 0,649 | 0,627 | 0,778 |
| Tabagisme | 0,801 | 0,856 | 0,925 | 0,959 |
| Consommation d'alcool | 0,566 | 0,721 | 0,563 | 0,632 |
| Consommation de drogues | 0,593 | 0,677 | 0,588 | 0,691 |
| Domicile : Oui | 0,242 | 0,535 | 0,135 | 0,474 |
| Domicile : Non | 0,474 | 0,439 | 0,667 | 0,667 |
| Activité physique : Oui | 0,373 | 0,522 | 0,373 | 0,513 |
| Activité physique : Non | 0,548 | 0,451 | 0,435 | 0,661 |
| Revenu | 0,333 | 0,333 | 0,933 | 0,963 |
| Niveau d'éducation | 0,167 | 0,380 | 0,444 | 0,440 |
| Origine ethnique | 0,143 | 0,247 | 0,100 | 0,458 |
| **Entités pour les relations** | | | | |
| StatusTime | 0,521 | 0,646 | 0,375 | 0,558 |
| Quantité | 0,782 | 0,871 | 0,845 | 0,891 |
| Durée | 0,096 | 0,369 | 0,417 | 0,742 |
| Fréquence | 0,450 | 0,679 | 0,230 | 0,622 |
| Antécédent | 0,479 | 0,628 | 0,506 | 0,660 |
| Type | 0,924 | 0,918 | 0,889 | 0,897 |

Tableau 3.22 : Accords inter-annotateurs pour chaque entité lors des deux premières phases d'annotation.

annotateurs ; (2) une seconde phase d'annotation commune entre Pacôme et moi-même sur 200 documents pour vérifier les modifications apportées aux règles d'annotation suite à la première phase ; (3) une phase d'annotation finale selon le guide d'annotation où chaque annotateur procède indépendamment.

Pour les entités, nous avons utilisé la librairie bratiia[1] pour calculer la F-mesure entre chaque paire d'annotateurs, à la fois au niveau de l'annotation et au niveau mot. Les accords présentés dans le tableau 3.22 sont les moyennes de toutes les paires d'annotateurs lorsqu'il y a plusieurs paires. Pour les relations, les accords inter-annotateurs ont également été calculés avec la F-mesure.

Les moyennes globales des accords inter-annotateurs pour chaque phase d'annotation sont données dans le tableau 3.23. Lors de la première phase d'annotation, la moyenne d'accord observé au niveau mot sur les entités était de 0,689 et a été amélioré à 0,725 lors de la seconde phase d'annotation. Le même phénomène est observé sur les relations avec une F-mesure de 0,795 lors de la première phase d'annotation, puis une F-mesure de 0,829 lors de la seconde phase.

| | Entités - Annotation F-mesure moyenne | Entités - Mot F-mesure moyenne | Relations F-mesure moyenne |
|---|---|---|---|
| Phase 1 | 0,564 | 0,689 | 0,795 |
| Phase 2 | 0,552 | 0,725 | 0,829 |

TABLEAU 3.23 : Moyenne globale des accords inter-annotateurs pour chaque phase d'annotation. Toutes les entités sont incluses dans la moyenne d'accord, les entités correspondant aux déterminants sociaux de santé, ainsi que les entités propres aux relations.

On observe, dans le tableau 3.22, une hausse d'accord entre les deux phases d'annotation pour la majorité des entités. La même tendance est observée pour les relations dans le tableau 3.24.

La troisième phase d'annotation a été réalisé par Pâcome et moi-même où nous nous sommes partagés 1 400 sections « Mode de vie » restantes à annoter. Les tableaux 3.25 et 3.26 rassemblent les fréquences d'apparition pour chaque type d'entités et relations dans les 1 700 sections annotées.

Nous observons que certains DSS sont très présents dans les comptes rendus, comme la descendance, le statut matrimonial, le statut d'emploi et la profession. D'autres déterminants sont beaucoup moins renseignés tels que le niveau d'éducation, le revenu ou l'origine ethnique malgré leur influence sur les inégalités sociales de santé.

1. `https://github.com/kldtz/bratiaa`

| | **Phase 1 3 annotateurs** | **Phase 2 2 annotateurs** |
|---|---|---|
| | *F-mesure* | *F-mesure* |
| **Relations** | | |
| Status | 0,873 | 0,902 |
| Quantité | 0,901 | 0,909 |
| Durée | 0,524 | 0,72 |
| Fréquence | 0,855 | 0,846 |
| Antécédent | 0,70 | 0,691 |
| Type | 0,917 | 0,907 |

TABLEAU 3.24 : Accords inter-annotateurs pour chaque relation lors des deux premières phases d'annotation.

| **Relation** | **Fréquence** |
|---|---|
| Antécédent | 362 |
| Durée | 77 |
| Fréquence | 783 |
| Quantité | 1 637 |
| Status | 1 665 |
| Type | 1 513 |

TABLEAU 3.25 : Fréquence par type de relation sur les 1 700 sections « Mode de vie ».

## 3.4 Conclusion

La majorité des systèmes de TAL actuels reposent sur un apprentissage supervisé qui va de pair avec la disponibilité des ressources textuelles annotées. Dans les domaines spécialisés, comme le domaine médical, les ressources sont rares dans les langues autres que l'anglais et cela entrave le développement de telles approches.

Dans ce chapitre, nous avons présenté les deux corpus cliniques annotés pendant cette thèse. Les deux premiers corpus décrits, **GAVROCHE** et déterminants sociaux de santé, sont composés de données cliniques issues de l'entrepôt de données de santé du CHU de Nantes. Le corpus dédié au projet **GAVROCHE** est constitué de comptes rendus hospitaliers entiers annotés avec des informations cliniques afin de phénotyper les patients hospitalisés pour insuffisance cardiaque aiguë. Le corpus sur les déterminants sociaux de santé est composé de paragraphes issus des sections « Mode de vie » des comptes rendus hospitaliers et annoté avec des 13 déterminants sociaux de santé. Ces deux corpus ont pour but de développer des systèmes de TAL permettant l'extraction automatique

| **Entités** | **Fréquence** |
|---|---|
| Activité physique : Oui | 208 |
| Activité physique : Non | 55 |
| Conditions de vie : Cohabitation | 600 |
| Conditions de vie : Seul | 281 |
| Consommation d'alcool | 697 |
| Consommation de drogues | 109 |
| Dernière profession exercée | 1 068 |
| Descendance : Oui | 1 172 |
| Descendance : Non | 136 |
| Domicile : Non | 22 |
| Domicile : Oui | 971 |
| Niveau d'éducation | 81 |
| Origine ethnique | 103 |
| Profession | 1 183 |
| Revenu | 40 |
| Statut matrimonial : Célibataire | 73 |
| Statut matrimonial : Divorcé | 97 |
| Statut matrimonial : Marié | 885 |
| Statut matrimonial : Veuf | 94 |
| Statut d'emploi : Actif | 495 |
| Statut d'emploi : Autre | 117 |
| Statut d'emploi : Etudiant | 48 |
| Statut d'emploi : Retraité | 441 |
| Statut d'emploi : Sans emploi | 162 |
| Tabagisme | 887 |
| **Entités pour les relations** | **Fréquence** |
| Antécédent | 360 |
| Durée | 78 |
| Fréquence | 781 |
| Quantité | 1 644 |
| StatusTime | 1 597 |
| Type | 1 524 |

Tableau 3.26 : Fréquence par type d'entité sur les 1 700 sections « Mode de vie ».

d'informations médicales pertinentes dans les CRH pour une réutilisation systématique dans la recherche clinique.

Concernant la publication des corpus, le corpus du projet **GAVROCHE** est composé de données cliniques sensibles et nous n'avons pas de solutions à l'heure actuelle pour

anonymiser le corpus afin de le rendre public. Pour le corpus sur les déterminants sociaux de santé, la diffusion du corpus est une piste explorée puisqu'il n'est pas composé de comptes rendus intégraux, mais de paragraphes issus de ces comptes rendus. De plus, nous pouvons encore descendre à une granularité plus fine et travailler sur des phrases isolées extraites de ces paragraphes. Les phrases isolées sont ciblées et portent uniquement sur les déterminants sociaux de santé avec des informations facilement partagées par plusieurs individus.

Chapitre 4

# Adaptation et évaluation de modèles de langue pré-entraînés pour le domaine biomédical français

Comme nous l'avons décrit dans le chapitre 2.4, les modèles de langue pré-entraînés sont devenus fondamentaux dans la conception de systèmes spécifiques aux tâches de TAL (Peters et al., 2018; Devlin et al., 2019). Depuis la parution de l'architecture *Transformer* et des premiers modèles de langue, des versions monolingues ou spécialisées de ces modèles ont vu le jour. Pour le domaine médical, dû à une plus grande disponibilité des données en anglais, seuls des modèles dans cette langue sont disponibles.

Dans ce chapitre, nous présentons plusieurs contributions de cette thèse sur l'adaptation de modèles de langue français aux domaines biomédical et clinique. Dans la section 4.1, nous introduisons DrBenchmark, un benchmark pour l'évaluation des modèles de langue pré-entraînés qui rassemble, à notre connaissance, la majorité des corpus biomédicaux français existant. Nous présentons ensuite, dans la section 4.2, une étude évaluant les stratégies de pré-entraînement de modèles BERT pour le domaine médical ainsi que l'impact des sources de données, en comparant les données biomédicales issues du web et les données cliniques issues du soin de patients. Grâce à cette étude, nous introduisons deux familles de modèles de langue pour le français : DrBERT et ChuBERT. Dans la section 4.3, nous présentons l'adaptation des modèles Longformer en comparant de nouveau les stratégies de pré-entraînement et en introduisant le modèle DrLongformer. Enfin, dans la section 4.4, nous discutons de l'impact écologique des travaux présentés.

## 4.1 DrBenchmark : un benchmark français de corpus biomédicaux

L'émergence des modèles de langues pré-entraînés sur de grands corpus de données textuelles a permis une application de ces modèles à un large éventail de tâches du TAL. L'évaluation des modèles et des approches proposées est une étape essentielle pour vérifier leur qualité et leurs performances. Dans le contexte des modèles de langue, cette validation consiste généralement à évaluer leurs performances sur des tâches cibles spécifiques. Le processus de sélection des tâches est crucial, car les performances des modèles peuvent varier en fonction des tâches choisies. Un modèle avec de bons résultats dans un contexte pourrait donner des résultats décevants dans un autre. Pour répondre à ce problème et valider la généralisabilité des modèles, des benchmarks d'évaluation ont vu le jour ces dernières années, englobant généralement divers ensembles de tâches. Ainsi, la disponibilité des benchmarks d'évaluation joue un rôle essentiel dans l'amélioration des modèles en facilitant les comparaisons équitables entre eux. L'un des premiers benchmarks, GLUE (Wang et al., 2018a), se concentre sur les tâches générales de compréhension de la langue anglaise plutôt que d'être spécifique à un domaine particulier. Ce benchmark a introduit neuf tâches, incluant la classification de textes (acceptabilité linguistique, analyse de sentiment, etc.), l'analyse sémantique (vérification de paraphrase, similarité de phrases, etc.), les systèmes de question-réponse (QA), la détection de coréférence, et l'inférence en langage naturel (NLI). Suivant la même approche, l'équivalent français de GLUE, connu sous le nom de FLUE (Le et al., 2020), comprend sept tâches générales en français telles que la classification de textes, la paraphrase, la NLI, l'analyse syntaxique et la désambiguïsation du sens des mots.

Bien que de nombreux benchmarks existent pour des tâches générales dans plusieurs langues, le biomédical reste un domaine avec relativement peu de benchmarks proposés. L'anglais et le chinois sont les plus majoritaires, facilitant ainsi la disponibilité de nombreux modèles biomédicaux dans ces deux langues. Par exemple, des plateformes comme BLURB (Gu et al., 2021) et BLUE (Peng et al., 2019a) offrent des benchmarks pour l'anglais, tandis que CBLUE (Zhang et al., 2022) est destiné à la langue chinoise. Ces benchmarks rassemblent des tâches identiques à celles du domaine général, telles que la reconnaissance d'entités nommées, l'extraction d'informations et de relations, la similarité entre les phrases, la classification de textes, les systèmes de question-réponse et l'inférence en langage naturel.

Dans cette section, nous introduisons DrBenchmark, un benchmark constitué de 20 tâches biomédicales françaises. L'objectif de ce benchmark est de rassembler les corpus du domaine biomédical français et de proposer un format d'évaluation permettant facilement de reproduire les expériences et de comparer équitablement des modèles. DrBenchmark est également conçu pour être facilement étendu à de nouvelles tâches. Avant de détailler ce benchmark, nous présentons tout d'abord l'ensemble des corpus et des tâches décrits dans la littérature. Puis, nous donnons la liste des tâches qui ont été rassemblées pour constituer DrBenchmark ainsi que les choix faits pour chaque tâche. Nous nous servons de ce benchmark comme base d'évaluation pour tous les modèles que nous introduisons dans les sections suivantes.

### 4.1.1 Corpus biomédicaux et cliniques français

Les corpus disponibles en français sont beaucoup moins nombreux que les corpus en langue anglaise. Dans le framework BigBIO (Fries et al., 2022), rassemblant plus de 126 corpus annotés, 83 % des corpus sont en anglais contre seulement 2,9 % pour le français. Les sources de données utilisées pour construire les corpus français sont moins variées que pour les corpus anglais, avec une surreprésentation du domaine biomédical par rapport au domaine clinique. En effet, la majorité des corpus français s'appuie sur des données libres d'accès, généralement issus de la recherche scientifique, telles que les résumés d'articles, les articles scientifiques intégraux et les fiches de prescriptions des médicaments. Néanmoins, les documents issus de la recherche scientifique restent principalement rédigés en anglais et seulement un faible sous-ensemble est disponible dans d'autres langues. Pour le domaine clinique, hormis les cas cliniques offrant une alternative semblable à des données cliniques, il n'existe aucun réel corpus clinique en libre accès. Le tableau 4.1 rassemble une liste des corpus français disponibles dans la littérature que nous présentons ci-dessous par source de données (biomédical, clinique ou mixte) puis par ordre chronologique.

#### 4.1.1.1 Corpus biomédicaux

**CRTT-MED** (Maniez, 2011)[1] est un corpus composé d'articles scientifiques du domaine médical issus des revues disponibles sur la base de données *Science Direct*. Ce corpus est composé 663 documents annotés automatiquement pour l'étiquetage morpho-syntaxique.

1. `https://perso.univ-lyon2.fr/~maniezf/Corpus/Corpus_medical_FR_CRTT.htm`

| Corpus | Tâche - Sous-ensemble | Source de données | Nombre | Licence |
|---|---|---|---|---|
| CAS corpus | Étiquetage morphosyntaxique | Cas cliniques | 7 580 | DUA |
| | CLS Multi-classes | | | |
| | REN - Incertitude | | | |
| | REN - Négation | | | |
| CLISTER | Similarité sémantique textuelle | Cas cliniques | 1 000 | DUA |
| CRTT-MED | Étiquetage morphosyntaxique | Articles de Science Direct | 663 | - |
| DEFT-2019 | REN | Cas cliniques | 717 | DUA |
| DEFT-2020 | Similarité sémantique textuelle | Cas cliniques, encyclopédies & médicaments | 1 010 | DUA |
| | CLS Multi-classes | | 1 102 | |
| DEFT-2021 | CLS Multi-étiquettes | Cas cliniques | 275 | DUA |
| | REN | | | |
| DiaMed | CLS Multi-classes | Cas cliniques | 739 | CC BY-SA 4.0 |
| E3C | REN - Clinique | Cas cliniques | 1 615 | CC BY-NC |
| | REN - Temporalité | | | |
| ESSAI | Étiquetage morphosyntaxique | Protocoles d'essais cliniques | 7 247 | DUA |
| | CLS Multi-classes | | 7 247 | |
| | REN - Incertitude | | 6 601 | |
| | REN - Négation | | 7 247 | |
| FrenchMedMCQA | Questions-réponses | Examens de pharmacie | 3 105 | Apache 2.0 |
| | CLS Multi-classes | | 3 105 | |
| Mantra-GSC | REN - EMEA | Résumés et titres d'articles biomédicaux, médicaments & brevets | 100 | CC BY 4.0 |
| | REN - Medline | | 100 | |
| | REN - Patents | | 50 | |
| MorFITT | CLS Multi-étiquettes | Résumés d'articles biomédicaux | 3 624 | CC BY-SA 4.0 |
| PxCorpus | REN | Transcriptions de prescriptions de médicaments | 1 981 | CC BY 4.0 |
| | CLS Multi-classes | | | |
| QUAERO | REN - EMEA | Notices de médicaments & titres d'articles biomédicaux | 38 | GFDL 1.3 |
| | REN - MEDLINE | | 2 498 | |

Tableau 4.1 : Synthèse des corpus français existants. Le nombre donné correspond au nombre de documents ou d'exemples dans le corpus. REN = Reconnaissance d'entités nommées. CLS = Classification

**QUAERO French Medical Corpus** (Névéol et al., 2014) est un corpus biomédical de reconnaissance d'entités nommées annoté avec dix catégories d'entités correspondant aux groupes sémantiques UMLS (Bodenreider, 2004) (plus de détails dans l'annexe 6.4). Le corpus couvre deux genres de textes, comprenant un total de 103 056 mots provenant d'EMEA ou de MEDLINE. Au total, 26 409 annotations d'entités ont été mises en correspondance avec 5 797 concepts UMLS uniques. Ce corpus inclut des annotations imbriquées, c'est-à-dire que plusieurs entités peuvent être associées à un même mot.

**Mantra-GSC** (Kors et al., 2015) est un corpus parallèle multilingue annoté pour la reconnaissance d'entités nommées biomédicales. Ce corpus couvre cinq langues : allemand, anglais, espagnol, français et néerlandais et est composé de trois sources de données : (1) des titres d'articles issus de Medline ; (2) des documents de l'agence européenne des

médicaments (*European Medicines Agency - EMEA*) et (3) des brevets de l'Office européen des brevets (*European Patent Office - EPO*). Chaque document récolté dans chacune des sources est décomposé en une ou plusieurs portions de texte. La partie française du corpus est composée de 100 portions de texte issues de titres d'articles, 100 portions de texte issues du corpus de médicament EMEA et 50 portions de texte issues des brevets. Deux schémas d'annotation sont utilisés : Medline (11 classes), EMEA et Patents (10 classes). La liste détaillée des classes pour chaque schéma est présentée dans l'annexe 6.7.

**E3C** (Magnini et al., 2020) est un corpus multilingue composé principalement de cas cliniques annotés pour la tâche de reconnaissance d'entités nommées. Ce corpus couvre cinq langues : anglais, basque, espagnol, français et italien. Selon la langue, la nature des données peut varier, allant de cas cliniques à des résumés d'articles scientifiques. Concernant la partie française du corpus, elle est composée de 1 615 cas cliniques et d'environ 8 000 notices de médicaments. Le corpus se compose de deux types d'annotations : (i) entités cliniques (par exemple, les pathologies), (ii) informations temporelles et factuelles (par exemple, les événements). La liste détaillée des entités pour chaque tâche est présentée dans l'annexe 6.5.

**PxCorpus** (Kocabiyikoglu et al., 2022) est un corpus de compréhension du langage parlé (*spoken language understanding*) dans le domaine des prescriptions de médicaments. Il comprend quatre heures de dialogues transcrits portant sur les prescriptions de médicaments en français. Le corpus a été créé par le biais d'une expérience impliquant 55 participants, à la fois experts et non-experts. L'ensemble de données comprend 1 981 enregistrements, dont 38 % proviennent de non-experts, 25 % de médecins et 36 % de praticiens médicaux. Les enregistrements ont été transcrits manuellement et annotés sémantiquement. La première tâche consiste à classer les énoncés textuels dans l'une des quatre classes d'intention. La seconde tâche porte sur la reconnaissance d'entités nommées dans laquelle chaque mot d'une séquence est classé dans l'une des 38 classes (plus de détails sur les classes dans l'annexe 6.10).

**FrenchMedMCQA** (Labrak et al., 2022) contient des questions à choix multiples provenant du concours d'internat en pharmacie. Les questions et les réponses associées ont été collectées sur le site web remede [2]. Ce site a été construit autour d'une communauté

2. `http://www.remede.org/internat/pharmacie/qcm-internat.html`

liée au domaine de la santé (médecine, pharmacie, odontologie, etc.), offrant de multiples informations (actualités, offres d'emploi, forums...) à la fois pour les étudiants mais aussi pour les professionnels de ce secteur d'activité. Les questions et réponses ont été créées manuellement par des experts médicaux et utilisées lors des examens. Le jeu de données récolté est composé de 2 025 questions à réponses multiples et de 1 080 questions à réponse unique, soit un total de 3 105 questions. Chaque exemple du jeu de données contient un identifiant, une question, cinq options (étiquetées de A à E) et la (les) bonne(s) réponse(s). Le corpus est décliné en deux tâches. La première est une tâche de questions-réponses qui consiste à identifier les réponses correctes parmi cinq réponses proposées pour une question donnée. La seconde est une tâche de classification de textes où l'objectif est de trouver le nombre de réponses correctes (entre 1 et 5) pour une question donnée.

**MorFITT** (Labrak et al., 2023b) est un corpus de classification multi-étiquettes qui a été annoté avec 12 spécialités du domaine médical (plus de détails dans l'annexe 6.6), pour un total de 5 116 annotations. Le corpus se compose de 3 624 résumés d'articles scientifiques obtenus à partir de *PMC Open Access*.

#### 4.1.1.2 Corpus cliniques

**CAS corpus** (Grabar et al., 2018) comprend 7 580 cas cliniques annotés pour quatre tâches : (1) l'étiquetage morphosyntaxique (*part-of-speech tagging*) avec 31 classes, (2) la classification multi-classes qui consiste à classer les phrases dans `NEGATION`, `INCERTITUDE` ou les deux catégories, (3) la reconnaissance d'entités nommées (REN) avec le marqueur de `NEGATION` et sa portée et (4) la REN avec le marqueur d'`INCERTITUDE` et sa portée. Nous renvoyons à l'annexe 6.2 pour une description plus détaillée de chaque tâche.

**DEFT-2019** (Grabar et al., 2019) est un corpus de reconnaissance d'entités nommées annoté avec quatre entités, comme détaillé dans l'annexe 6.8. Ce corpus est constitué de 717 cas cliniques issus du corpus CAS et a fait l'objet de l'édition 2019 du défi fouilles de textes (DEFT).

**ESSAI** (Dalloux et al., 2021) contient 13 848 cas cliniques annotés pour quatre tâches : (1) l'étiquetage morphosyntaxique (*part-of-speech tagging*) avec 35 étiquettes, (2) la classification multi-classes qui consiste à classer les phrases dans `NEGATION`, `INCERTITUDE` ou les deux catégories, (3) la REN avec le marqueur de `NEGATION` et sa portée, et (4) la REN

avec le marqueur d'`INCERTITUDE` et sa portée, plus de détails pour chaque tâche dans l'annexe 6.3).

**DEFT-2021** (Grouin et al., 2021) est un sous-ensemble de 275 cas cliniques tirés de l'édition 2019 de DEFT. Ce corpus a été annoté manuellement pour deux tâches : (i) la classification multi-étiquettes et (ii) la REN. La tâche de classification multi-label se concentre sur l'identification du profil clinique du patient en fonction des maladies, des signes ou des symptômes mentionnés dans les cas cliniques. Le corpus est annoté avec 23 axes dérivés du chapitre C des *Medical Subject Headings* (MeSH). La deuxième tâche consiste à extraire des informations fines pour 13 entités, dont le détail est donné dans l'annexe 6.9.

**CLISTER** (Hiebel et al., 2022) est un corpus de similarité sémantique textuelle clinique de 1 000 paires de phrases. Chaque paire de phrases a été annotée manuellement par plusieurs humains en attribuant des scores de similarité allant de 0 à 5 à chaque paire. La moyenne des scores est ensuite calculée pour obtenir un nombre à virgule flottante représentant la similarité globale. L'objectif de ce corpus est de développer des modèles capables de prédire automatiquement un score de similarité proche du score de référence en se basant uniquement sur les deux phrases fournies.

**DiaMed** est un corpus que nous avons développé. DiaMed est composé de 739 cas cliniques annotés avec 22 catégories de la Classification Internationale des Maladies, 10ème Révision (CIM-10) pour une tâche de classification multi-classes. Pour simplifier l'annotation, seuls les grands chapitres de la CIM-10 ont été annotés. L'annotation a été réalisée par quatre annotateurs dont un neurologue (Dr Pacôme Constant dit Beaufils) et trois annotateurs experts en TAL biomédical (Oumaima El Khettari, Yanis Labrak et moi-même). Pour évaluer l'accord entre les annotateurs, plusieurs sessions d'annotation ont été menées où 15 documents étaient annotés par tous les annotateurs à chaque session. Des séances d'adjudication entre chaque session d'annotation ont permis de traiter les discordances entre les annotateurs et parvenir à un consensus pour les cas ambigus. Les mesures de Kappa de Cohen (introduite dans la section 3.2.3.1) et de Gwet AC1 (introduite dans la section 3.2.3.1) pour chaque session sont présentées dans le tableau 4.2. Après deux sessions d'annotation, un Kappa de Cohen de 0,714 a été mesuré et nous avons considéré ce score suffisant pour initier l'annotation du corpus entier.

| | Session 1 - 0 à 15 docs | | Session 2 - 15 à 30 docs | |
|---|---|---|---|---|
| ID Annotateur | Kappa Cohen | Gwet AC1 | Kappa Cohen | Gwet AC1 |
| Annotateurs 1 & 2 | 0,538 | 0,566 | 0,697 | 0,705 |
| Annotateurs 1 & 3 | 0,682 | 0,709 | 0,697 | 0,705 |
| Annotateurs 1 & 4 | 0,397 | 0,429 | 0,548 | 0,558 |
| Annotateurs 2 & 3 | 0,311 | 0,357 | 1,000 | 1,000 |
| Annotateurs 2 & 4 | 0,472 | 0,497 | 0,672 | 0,707 |
| Annotateurs 3 & 4 | 0,311 | 0,354 | 0,672 | 0,707 |
| Moyenne | 0,452 | 0,485 | 0,714 | 0,730 |

Tableau 4.2 : Accords inter-annotateurs pour le corpus DiaMed.

#### 4.1.1.3 Corpus mixtes

**DEFT-2020** (Cardon et al., 2020) est un corpus de cas cliniques introduit dans l'édition 2020 du défi fouille de textes (DEFT). Ce corpus est annoté pour deux tâches : (i) la similarité sémantique textuelle et (ii) la classification multi-classes. La première tâche vise à identifier le degré de similarité entre les paires de phrases, allant de 0 (les moins similaires) à 5 (les plus similaires). La seconde tâche consiste à identifier la phrase la plus proche d'une phrase source parmi trois phrases données.

### 4.1.2 Constitution de DrBenchmark

Le principal critère de sélection était d'avoir un ensemble de tâches diversifiées provenant de différentes sources de données et non pas de sélectionner seulement les jeux de données les plus utilisés. Finalement, parmi l'ensemble des corpus présentés dans la section 4.1.1, 12 corpus ont été sélectionnés pour constituer la première version de DrBenchmark, pour un total de 20 tâches. Seuls les corpus CRTT-MED, DEFT-2019 et les tâches sur la négation et l'incertitude des corpus CAS et ESSAI n'ont pas été retenus. Le corpus CRTT-MED est peu connu de la communauté et très conséquent en taille pour une tâche de POS tagging déjà traitée dans les corpus CAS et ESSAI. Les autres tâches n'ont pas été intégrées dans cette première version du benchmark car les résultats obtenus avec les modèles n'étaient pas stables. Ces corpus seront réadaptés pour être intégrés ultérieurement dans le benchmark. L'ensemble des tâches constituant le benchmark sont présentées dans le tableau 4.3.

| Corpus | Tâche | Métrique | Train | Validation | Test | Long. Moy Seq. | Long. Max. Seq. |
|---|---|---|---|---|---|---|---|
| CAS corpus | POS | SeqEval F1 | 5 306 | 758 | 1 516 | 23.07 | 136 |
| CLISTER | SST | MSE / Spearman | 499 | 101 | 400 | 25.21 | 103 |
| DEFT-2020 | SST | MSE / Spearman | 498 | 102 | 410 | 45.14 | 118 |
| | CLS | Weighted F1 | 460 | 112 | 530 | 90.19 | 254 |
| DEFT-2021 | CLS multi-étiquettes | Weighted F1 | 118 | 49 | 108 | 332.55 | 1 454 |
| | REN | SeqEval F1 | 2 153 | 793 | 1 766 | 23.62 | 509 |
| DiaMed | CLS | Weighted F1 | 509 | 76 | 154 | 329.92 | 1 379 |
| E3C | REN - Clinical | SeqEval F1 | 969 | 140 | 293 | 21.51 | 143 |
| | REN - Temporal | SeqEval F1 | 969 | 140 | 293 | 22.07 | 88 |
| ESSAI | POS | SeqEval F1 | 9 693 | 1 385 | 2 770 | 22.55 | 135 |
| FrenchMedMCQA | QA | Hamming / EMR | 2 171 | 312 | 622 | 44.80 | 174 |
| | CLS | Weighted F1 | 2 171 | 312 | 622 | 44.80 | 174 |
| Mantra-GSC | REN - EMEA | SeqEval F1 | 70 | 10 | 20 | 21.22 | 63 |
| | REN - Medline | SeqEval F1 | 70 | 10 | 20 | 10.03 | 21 |
| | REN - Patents | SeqEval F1 | 35 | 5 | 10 | 48.58 | 284 |
| MorFITT | CLS multi-étiquettes | Weighted F1 | 1 514 | 1 022 | 1 088 | 226.33 | 1 425 |
| PxCorpus | REN | SeqEval F1 | 1 386 | 198 | 397 | 11.33 | 48 |
| | CLS | Weighted F1 | 1 386 | 198 | 397 | 11.33 | 48 |
| QUAERO | REN - EMEA | SeqEval F1 | 429 | 389 | 348 | 29.90 | 282 |
| | REN - Medline | SeqEval F1 | 833 | 832 | 833 | 12.78 | 48 |

TABLEAU 4.3 : Descriptions des tâches incluses dans DrBenchmark. **POS** : *part-of-speech tagging*. **SST** : similarité sémantique textuelle. **CLS** : classification de séquences. **REN** : reconnaissance d'entités nommées. **QA** : *question answering*. **Long. Moy. Seq.** : Longueur moyenne des séquences. **Long. Max. Seq.** : Longueur maximum des séquences.

### 4.1.3 Formatage et métriques d'évaluation

Cette première version de DrBenchmark a pour objectif d'évaluer les modèles de langue français basés sur l'architecture BERT. Ces modèles étant limités à 512 *tokens* en entrée, les corpus ayant des séquences supérieures à 512 *tokens* ont été découpés en phrases. Ce traitement a été appliqué aux corpus E3C, DEFT-2021 et au sous-ensemble EMEA du corpus QUAERO. Les corpus ESSAI, CAS, E3C et Mantra-GSC n'étaient pas fournis avec des sous-ensembles d'entraînement, de validation et de test prédéfinis. Nous avons pris la décision de le diviser aléatoirement en 3 sous-ensembles de 70 %, 10 % et 20 % des données totales pour l'entraînement, la validation et le test respectivement. Pour assurer une comparaison équitable et cohérente entre les systèmes pour les tâches de reconnaissance d'entités nommées et d'étiquetage morphosyntaxique, nous avons converti les corpus au format IOB2 (*inside-outside-beginning)* et évaluons avec la métrique SeqEval (Nakayama, 2018). L'évaluation avec SeqEval sur un format IOB2 permet d'entraîner et d'évaluer les modèles à prédire uniquement l'étiquette du premier *token* de chaque mot. Cela fournit une évaluation agnostique au *tokenizer* et atténue toute corrélation entre les performances des modèles et leur *tokenizer*. Cependant, ce format ne permet pas d'avoir plusieurs étiquettes sur un même mot. Pour les corpus intégrant des annotations imbriquées comme QUAERO, nous avons suivi l'approche proposée par Touchent et al. (2023) et seules les

annotations de plus large granularité ont été conservées. Cette suppression des annotations se traduit par une perte moyenne de 6,06 % des annotations pour le sous-ensemble EMEA et 8,90 % pour Medline dans le corpus QUAERO. Concernant les autres métriques d'évaluation, les tâches de classification de séquences sont évaluées avec un score F1 pondéré, les tâches de similarité sémantique textuelle sont évaluées avec l'erreur quadratique moyenne (*mean squared error* - MSE) et la corrélation de Spearman, et la tâche de question-réponse est évaluée avec le score de Hamming et l'*exact match ratio* (EMR).

### 4.1.4 Utilisation du benchmark et reproductibilité

Pour faciliter l'adoption de DrBenchmark et assurer une cohérence dans les implémentations des scripts pour chaque corpus, nous avons repris le formatage de la bibliothèque Datasets de HuggingFace (Lhoest et al., 2021). Cette librairie intègre des chargeurs de données (*data loaders*) qui respectent des schémas normalisés et des sous-ensembles de données prédéfinies. Elle fournit également des scripts de pré-entraînement et d'évaluation pour chacune des tâches, en utilisant les librairies *Transformers* de HuggingFace (Wolf et al., 2020) et PyTorch (Paszke et al., 2019). Tous les scripts nécessaires à l'utilisation du benchmark sont disponibles sur Github[3], incluant les scripts bash permettant d'exécuter le benchmark sur divers environnements dont SLURM.

3. https://anonymous.4open.science/r/DrBenchmark-4F7E/

## 4.2 Adaptation de modèles BERT aux domaines biomédical et clinique français

Dans le chapitre 2, nous avons vu que plusieurs travaux ont été proposés pour l'adaptation des modèles de langue au domaine médical, notamment pour l'anglais. Cependant, aucune comparaison n'a été faite entre les modèles de langue pré-entraînés sur des données spécifiques au domaine provenant du web et ceux pré-entraînés sur des documents cliniques issus d'entrepôts de données de santé, dont la qualité peut être contrôlée.

Nous présentons dans cette section une adaptation des travaux présentés dans la publication de Labrak et al. (2023a). Ces travaux ont été réalisés en collaboration avec Yanis Labrak, doctorant au Laboratoire d'Informatique d'Avignon (LIA) et ses encadrants, Mickael Rouvier et Richard Dufour. Les contributions de ces travaux sont multiples. Premièrement, nous décrivons et diffusons librement DrBERT, les premiers modèles de langue pré-entraînés pour le domaine biomédical français, ainsi que le corpus NACHOS qui a permis leur entraînement. Nous proposons ensuite une étude évaluant les différentes stratégies de pré-entraînement de modèles de langue pour le domaine médical, en comparant le pré-entraînement de zéro et le pré-entraînement poursuivi. Nous évaluons également dans cette étude l'impact des sources de données, en comparant les modèles DrBERT, des modèles pré-entraînés sur des données biomédicales récoltées sur le web, avec ChuBERT, un modèle pré-entraîné sur des données cliniques issues de l'entrepôt de données du CHU de Nantes. Enfin, nous évaluons les modèles au niveau syntaxique et sémantique (classification multi-étiquettes, étiquetage morphosyntaxique, reconnaissance d'entités nommées, etc.) sur les tâches de DrBenchmark.

### 4.2.1 Données de pré-entraînements

Dans le domaine biomédical, les travaux précédents sur les modèles de langue pré-entraînés ont souligné l'importance d'adapter les sources de données utilisées pour leur pré-entraînement aux tâches cibles en aval (Gu et al., 2021). En raison de leur nature sensible, les données cliniques sont difficiles à obtenir et à utiliser pour le pré-entraînement de modèles puisque ceux-ci ne peuvent généralement pas être publiés. La collecte massive de données web liées à ce domaine semble être une solution pour pallier ce manque. Cependant, ces documents web varient en termes de qualité.

Nous avons collecté deux corpus pour pré-entraîner les modèles. Le tableau 4.4 donne

un aperçu général des deux corpus collectés. Les données publiques basées sur le web, détaillées dans la section 4.2.1.1, ont permis de constituer un corpus nommé NACHOS$_{large}$ contenant 7,4 Go de données. Le corpus privé, appelé NBDW$_{small}$, est décrit dans la section 4.2.1.2 et contient 4 Go de données. Afin de pouvoir comparer équitablement les modèles sur les deux sources de données différentes, nous avons extrait un sous-corpus de NACHOS (NACHOS$_{small}$) de la même taille que les données privées. Enfin, la section 4.2.1.3 décrit les prétraitements appliqués aux deux corpus.

| Corpus | Source de données | Taille | #mots | #phrases | Disponibilité |
|---|---|---|---|---|---|
| NACHOS$_{large}$ | Biomédical | 7,4 GB | 1,1 B | 54,2 M | Public |
| NACHOS$_{small}$ | Biomédical | 4 GB | 646 M | 25,3 M | Public |
| NBDW$_{small}$ | Clinique | 4 GB | 655 M | 43,1 M | Privé |
| NBDW$_{mixed}$ | Biomédical+Clinique | 4+4 GB | 1,3 B | 68,4 M | Privé |

Tableau 4.4 : Aperçu des corpus public (NACHOS) and privé (NBDW) collectés.

#### 4.2.1.1 Données publiques : NACHOS

NACHOS (opeN crAwled frenCh Healthcare cOrpuS) est un ensemble de données textuelles médicales françaises obtenu à partir d'une collecte massive sur le web de plusieurs sources textuelles ayant comme thématique le médical. Le corpus se compose de plus d'un milliard de mots, extraits de 24 sites web francophones institutionels garantissant une fiabilité des données. Il comprend un large éventail d'informations médicales : descriptions de maladies, fiches d'information sur les traitements et les médicaments, conseils généraux sur la santé, rapports de réunions scientifiques officielles, cas cliniques anonymisés, littérature scientifique, thèses, cours de médecine universitaire, ainsi que de nombreuses données obtenues à partir de sources textuelles brutes, de *scrapping web* et de Reconnaissance Optique de Caractères (OCR) comme présenté dans le tableau 4.5. Des heuristiques ont été utilisées pour découper les textes en phrases et filtrer les phrases courtes ou de mauvaise qualité comme celles obtenues par OCR. Ensuite, nous les classons par langue en utilisant notre propre classifieur entraîné sur les corpus multilingues Opus EMEA (Tiedemann and Nygaard, 2004) et MASSIVE (FitzGerald et al., 2022) afin de ne garder que les phrases en français. Pour la version 4 Go de NACHOS (NACHOS$_{small}$), nous avons mélangé l'ensemble du corpus et sélectionné au hasard 25,3 millions de phrases afin de maximiser l'homogénéité des sources de données.

| Nom de la source | # mots |
|---|---|
| HAL | 638 508 261 |
| Haute Autorité de Santé (HAS) | 113 394 539 |
| Drug leaflets | 74 770 229 |
| Medical Websites Scrapping | 64 904 334 |
| ANSES SAISINE | 51 372 932 |
| Public Drug Database (BDPM) | 48 302 695 |
| ISTEX | 44 124 422 |
| CRTT | 26 210 756 |
| WMT-16 | 10 282 494 |
| EMEA-V3 | 6 601 617 |
| Wikipedia Life Science French | 4 671 944 |
| ANSES RCP | 2 953 045 |
| Cerimes | 1 717 552 |
| LiSSa | 235 838 |
| DEFT-2020 | 231 396 |
| CLEAR | 225 898 |
| CNEDiMTS | 175 416 |
| QUAERO French Medical Corpus | 72 031 |
| ANSM Clinical Study Registry | 47 678 |
| ECDC | 44 482 |
| QualiScope | 12 718 |
| WMT-18-Medline | 7 673 |
| **Total** | **1 088 867 950** |

Tableau 4.5 : Sources des données constituant le corpus NACHOS.

#### 4.2.1.2 Données cliniques : CHU de Nantes

Le corpus privé, appelé Nantes Biomedical Data Warehouse (NBDW), a été obtenu en utilisant l'entrepôt de données du CHU de Nantes. Cet entrepôt de données comprend différentes dimensions de données relatives aux patients : socio-démographiques, prescriptions médicamenteuses et autres informations associées au séjour hospitalier (diagnostic, biologie, imagerie, etc.). L'autorisation de mise en œuvre et d'exploitation de l'entrepôt de données a été accordée en 2018 par la CNIL (Commission Nationale de l'Informatique et des Libertés) ; autorisation N°2129203. Pour ce travail, un échantillon de 1,7 millions de comptes-rendus hospitaliers désidentifiés a été sélectionné aléatoirement et extrait de l'entrepôt de données. Les comptes-rendus proviennent de différents services hospitaliers, tels que les urgences, la gynécologie et la cardiologie, comme décrit dans le tableau 4.6. Ce corpus contient 655 millions de mots, issus de 43,1 millions de phrases, pour une taille

totale d'environ 4 Go. Chacun des comptes-rendus a été découpé en phrases avec une moyenne de 15,26 mots par phrase. Toutes les phrases de tous les comptes-rendus ont ensuite été mélangées pour construire le corpus de pré-entraînement. Au total, ce corpus contient 655M de mots, issus de 43,1M de phrases, pour une taille totale d'environ 4 Go.

| **Spécialité médicale** | **# documents** | **# mots** |
|---|---|---|
| Autre | 474 588 | 192 832 792 |
| Urgences | 235 579 | 90 807 406 |
| Soins ambulatoires | 119 149 | 50 975 472 |
| Consultation | 95 135 | 38 335 804 |
| Gynécologie | 132 983 | 38 204 495 |
| Cardiologie | 29 633 | 22 654 583 |
| Oncologie | 45 603 | 22 587 869 |
| Gastro-entérologie | 46 600 | 21 340 794 |
| Chirurgie orthopédique | 82 084 | 18 983 791 |
| Hématologie | 41 776 | 18 285 983 |
| Soins intensifs et Réanimation | 20 819 | 16 472 785 |
| ORL | 69 343 | 16 131 214 |
| Dermatologie | 51 804 | 15 035 412 |
| Rhumatologie | 31 527 | 14 647 543 |
| Urologie | 51 535 | 14 272 231 |
| Chirurgie du côlon et du rectum | 45 987 | 13 334 550 |
| Médecine interne | 23 904 | 13 282 253 |
| Psychiatrie | 26 628 | 12 496 503 |
| Neurochirurgie | 34 481 | 10 360 533 |
| Néphrologie | 19 171 | 9 548 533 |
| Ophthalmologie | 19 700 | 4 464 515 |
| **Total** | **1 698 029** | **655 055 061** |

Tableau 4.6 : Sources des données constituant le corpus NBDW classé par taille en nombre de mots.

#### 4.2.1.3 Prétraitements des corpus

Les données textuelles des deux corpus ont été segmentées en tokens en utilisant un tokenizer exploitant l'algorithme SentencePiece (Kudo and Richardson, 2018) avec une taille de vocabulaire de 32 000 tokens pour les modèles pré-entraîné à partir de zéro (voir section 4.2.2.2). Pour chacun des corpus, le *tokenizer* est entraîné à partir de toutes les phrases du jeu de données de pré-entraînement. Pour les modèles dont le pré-entraînement

a été poursuivi, les deux corpus ont été segmentés en utilisant le *tokenizer* du modèle initial.

### 4.2.2 Pré-entraînements des modèles

Dans cette section, nous décrivons les modalités de pré-entraînement de nos modèles selon deux points de vue : 1) l'influence des données utilisées (taille et nature), et 2) les stratégies de pré-entraînement des modèles. Ces deux points de vue sont respectivement détaillés dans les sections 4.2.2.1 et 4.2.2.2.

#### 4.2.2.1 Influence des données

L'une des problèmatiques consiste à identifier la quantité de données nécessaires pour créer un modèle performant et capable de rivaliser avec les modèles pré-entraînés sur des domaines généraux. Des études récentes, comme celles de Martin et al. (2020) et Zhang et al. (2020), discutent de l'impact de la taille des données de pré-entraînement sur la performance des modèles et montrent que certaines tâches bénéficient d'une quantité moindre de données. Dans le domaine médical, aucune étude n'a été menée pour comparer l'impact de la variation de la quantité de données sur le pré-entraînement ou pour évaluer l'impact de la qualité des données en fonction de leur source de collecte.

Nous proposons donc d'évaluer l'impact des différentes sources de données en comparant $\text{NACHOS}_{small}$ et $\text{NBDW}_{small}$ entre elles comme décrit dans la section 4.2.1. De plus, nous proposons de comparer les résultats obtenus avec ceux d'un modèle pré-entraîné sur une quantité de données plus grande ($\text{NACHOS}_{large}$) afin d'étudier si le fait de disposer de presque deux fois plus de données permet d'améliorer les performances. Pour finir, nous évaluons une combinaison des corpus public ($\text{NACHOS}_{small}$) et privé ($\text{NBDW}_{small}$) pour un total de 8 Go ($\text{NBDW}_{mixed}$), afin de vérifier si la combinaison de données de sources différentes et de qualité variable permettent des représentations complémentaires.

#### 4.2.2.2 Stratégie de pré-entraînement

En plus de l'analyse sur la taille et les sources des données, nous cherchons également à évaluer trois stratégies de pré-entraînement des modèles de langue pour le domaine médical :

- Pré-entraîner un modèle à partir de zéro, incluant le tokenizer à partir des données de pré-entraînement.

- Poursuivre le pré-entraînement d'un modèle de langue général pour le français, ici `CamemBERT`, sur nos données du domaine médical, tout en conservant le tokenizer initial (*i.e.* celui de `CamemBERT`).
- Poursuivre le pré-entraînement d'un modèle de langue spécifique au domaine médical anglais, ici `PubMedBERT`, sur nos données en français, en conservant le tokenizer initial.

Concernant la dernière stratégie, l'objectif est d'évaluer le pré-entraînement adaptatif à phases multiples avec différents modèles de départ. Plus précisément, nous souhaitons vérifier s'il est plus avantageux de continuer le pré-entraînement d'un modèle déjà adapté au domaine médical anglais plutôt que de continuer le pré-entraînement d'un modèle général français. En effet, ces deux langues morphologiquement proches partagent une terminologie de spécialité commune.

| Modèle | Architecture | Stratégie de pré-entraînement | Corpus |
|---|---|---|---|
| `DrBERT-7GB` | RoBERTa | De zéro | $\text{NACHOS}_{large}$ |
| `DrBERT-4GB` | RoBERTa | De zéro | $\text{NACHOS}_{small}$ |
| `ChuBERT-4GB` | RoBERTa | De zéro | $\text{NBDW}_{small}$ |
| `ChuBERT-Mixed` | RoBERTa | De zéro | $\text{NBDW}_{mixed}$ |
| `DrBERT-CP-CamemBERT` | RoBERTa | Poursuivi (`CamemBERT`) | $\text{NACHOS}_{small}$ |
| `DrBERT-CP` | BERT | Poursuivi (`PubMedBERT`) | $\text{NACHOS}_{small}$ |
| `ChuBERT-CP-CamemBERT` | RoBERTa | Poursuivi (`CamemBERT`) | $\text{NBDW}_{small}$ |

Tableau 4.7 : Liste des configurations des modèles pré-entraînés.

Le tableau 4.7 résume toutes les configurations évaluées, intégrant à la fois l'étude de la taille des données et les stratégies de pré-entraînement.

**Architecture du modèle** Tous les modèles pré-entraînés à partir de zéro reprennent la configuration `CamemBERT` $_{\texttt{base}}$, identique à l'architecture `RoBERTa` $_{\texttt{base}}$ (12 couches, 768 dimensions cachées, 12 têtes d'attention, 110M de paramètres).

**Modélisation de la langue** Nous pré-entraînons les modèles sur la tâche de *Masked Language Modeling* (MLM) en utilisant la librairie HuggingFace (Wolf et al., 2019). Dans les modèles `BERT` et `RoBERTa` (y compris `CamemBERT`), 15 % des tokens sont choisis aléatoirement. Parmi ces tokens sélectionnés, 80 % sont remplacés par le token `<mask>`, 10 % restent inchangés et 10 % sont aléatoirement remplacés par un token du vocabulaire. Nous conservons cette probabilité de masquage de 15 % pour le pré-entraînement des modèles.

**Optimisation et pré-entraînement** Les modèles sont optimisés pendant 80k étapes (*steps*) avec des lots (*batch size*) de 4 096 séquences, chaque séquence étant remplie de 512 *tokens*, permettant de traiter 2,1M de *tokens* par étape. Le taux d'apprentissage (*learning rate*) est progressivement augmenté de manière linéaire pendant 10k étapes, passant de zéro au taux d'apprentissage initial de $5\times10^{-5}$. Les modèles sont entraînés sur 128 GPUs Nvidia V100 de 32 GB pendant 20 heures sur le supercalculateur Jean Zay. Nous utilisons l'entraînement à précision mixte (FP16) (Micikevicius et al., 2017) pour réduire l'empreinte mémoire, ce qui nous permet d'augmenter la taille du lot à 32 séquences sur chaque GPU.

### 4.2.3 Évaluation des modèles sur DrBenchmark

La récente émergence de modèles pré-entraînés pour le domaine biomédical français, avec les modèles `AliBERT` (Berhe et al., 2023), `CamemBERT-Bio` (Touchent et al., 2023) ou les modèles `DrBERT` (Labrak et al., 2023a) présentés dans la section 4.2.2.2, nous pousse à nous interroger sur l'évaluation de ces modèles. En effet, bien que les mêmes corpus soient souvent utilisés pour évaluer ces modèles, il est difficile de comparer directement les résultats en raison des nombreux paramètres susceptibles de changer lors de l'évaluation : *batch size*, *learning rate*, nombre d'itérations sur le corpus, manière d'aborder la tâche ou encore le découpage en sous-ensembles d'apprentissage, de validation et de test.

Nous proposons d'évaluer sur DrBenchmark les modèles présentés dans la section 4.2.2.2 ainsi que six modèles pré-entraînés état de l'art basés sur l'architecture BERT (Devlin et al., 2019) et disponibles en ligne : `CamemBERT` (Martin et al., 2020), `CamemBERTa` (Antoun et al., 2023a), `FlauBERT` (Le et al., 2020), `XLM-RoBERTa` (Conneau et al., 2020), `PubMedBERT` (Gu et al., 2021), `CamemBERT-bio` (Touchent et al., 2023). Pour simplifier la présentation des résultats, les modèles `DrBERT-CP-CamemBERT` et `ChuBERT-CP-CamemBERT` ne sont pas inclus dans l'évaluation et non utilisons `CamemBERT-bio` comme référence pour la stratégie de pré-entraînement poursuivi à partir de `CamemBERT`. Les performances des modèles `DrBERT-CP-CamemBERT` et `ChuBERT-CP-CamemBERT` après affinage sur les tâches sont instables et systématiquement en dessous de celles de `CamemBERT-bio`.

#### 4.2.3.1 Modèles de langue pré-entraînés

Le tableau 4.8 donne un aperçu des caractéristiques des modèles que nous décrivons ci-dessous.

| | Modèle | Tokenizer | Vocabulaire | Pré-entraînement | Corpus | Taille du corpus |
|---|---|---|---|---|---|---|
| Généraliste français | CamemBERTa | SentencePiece 32K | CCNET | De zéro | CCNET | 4 GB |
| | CamemBERT | SentencePiece 32K | OSCAR | De zéro | OSCAR | 138 GB |
| | FlauBERT | BPE 50K | Wiki + Web crawl | De zéro | Wiki + Web crawl | 71 GB |
| Biomédical français | DrBERT-7GB | SentencePiece 32K | NACHOS | De zéro | NACHOS | 7,4 GB |
| | DrBERT-4GB | SentencePiece 32K | NACHOS | De zéro | NACHOS | 4 GB |
| | DrBERT-CP | WordPiece 30K | PubMed | Poursuivi | PubMed + NACHOS | 21 + 4 GB |
| | CamemBERT-bio | SentencePiece 32K | OSCAR | Poursuivi | OSCAR + biomed-fr | 138 + 2,7 GB |
| Clinique français | ChuBERT-4GB | SentencePiece 32K | NBDW | De zéro | NBDW | 4 GB |
| | ChuBERT-Mixed | SentencePiece 32K | NBDW + NACHOS | De zéro | NBDW + NACHOS | 8 GB |
| Biomédical anglais | PubMedBERT | WordPiece 30K | PubMed | De zéro | PubMed | 21 GB |
| Généraliste multilingue | XLM-RoBERTa | WordPiece 30K | CC-100 | De zéro | CC-100 | 2.5 TB |

Tableau 4.8 : Aperçu des caractéristiques des modèles évalués.

**CamemBERT** (Martin et al., 2020) est un modèle français basé sur l'architecture RoBERTa et pré-entraîné de zéro sur la partie française de 138 GB du corpus `OSCAR` (Ortiz Suárez et al., 2019).

**CamemBERTa** (Antoun et al., 2023b) est un modèle reprenant l'architecture DeBERTaV3 (He et al., 2023) et pré-entraîné de zéro sur $\sim$ 30 % de la partie français du corpus `CCNET` (Wenzek et al., 2020) utilisée pour $\texttt{CamemBERT}_{CCNET}$.

**FlauBERT** (Le et al., 2020) est un modèle BERT français pré-entraîné de zéro avec un échantillon de 71 GB provenant du Common Crawl et du Wikipedia français.

**XLM-RoBERTa** (Conneau et al., 2020) est un modèle multilingue basé sur l'architecture RoBERTa pré-entraîné sur 116 langues, incluant le français avec 2,5 TB du corpus Common Crawl.

**PubMedBERT** (Gu et al., 2021) est un modèle BERT biomédical pré-entraîné de zéro sur 3,1 milliards de mots issus de PubMed (21 GB).

**CamemBERT-bio** (Touchent et al., 2023) est un modèle biomédical français basé sur le modèle $\texttt{CamemBERT}_{OSCAR-138GB}$ dont le pré-entraînement a été poursuivi sur le corpus `biomed-fr` composé de 413 millions de mots (2.7 GB) récoltés sur le web.

#### 4.2.3.2 Résultats

Nous proposons d'analyser les résultats sous deux angles. D'abord, dans la section 4.2.3.3, nous comparons les résultats obtenus par les modèles sur les tâches de DrBenchmark, nous permettant ainsi de positionner ces modèles état de l'art. Puis, dans la section 4.2.3.4, nous analysons la tokenisation des mots obtenue avec les modèles.

#### 4.2.3.3 Comparaison des performances des modèles

Les résultats des modèles sont présentés type de tâche dans les tableaux 4.9 pour la tâche d'étiquetage morphosyntaxique, 4.10 pour les tâches de reconnaissance d'entités nommées, 4.11 pour les tâches de classification et 4.12 pour les tâches de similarité sémantique textuelle et de *question answering*. Dans l'ensemble, nous observons qu'aucun modèle ne surpasse tous les autres et que la plupart des modèles obtiennent les meilleurs résultats sur au moins l'une des tâches.

| **Domaine** | **Modèle** | **CAS** | **ESSAI** |
|---|---|---|---|
| Général français | `CamemBERT` | 95,49** | 97,64** |
| | `CamemBERTa` | 96,90** | 98,27** |
| | `FlauBERT` | 95,61** | 95,86* |
| Biomédical français | `DrBERT-7GB` | <u>97,37</u> | 98,57 |
| | `DrBERT-4GB` | **97,40** | **98,61** |
| | `DrBERT-CP` | 96,87** | 98,19** |
| | `CamemBERT-bio` | 95,72** | 97,70** |
| Clinique français | `ChuBERT-4GB` | 97,05** | 98,46** |
| | `ChuBERT-Mixed` | 97,30 | <u>98,60</u> |
| Biomédical anglais | `PubMedBERT` | 95,38** | 97,67** |
| Général multilingue | `XLM-RoBERTa` | 97,32 | 98,50** |

Tableau 4.9 : Performances moyennes sur quatre exécutions pour les tâches d'étiquetage morphosyntaxique. Le meilleur modèle est en gras et le second est souligné. La significativité statistique est calculée en utilisant le test de Student : * signifie $p < 0{,}05$, ** signifie $p < 0{,}01$.

**Général *vs.* Spécialisé.** La nature des données semble avoir une influence sur les performances. Les modèles généralistes tels que `CamemBERT`, `CamemBERTa`, `FlauBERT` et `XLM-RoBERTa` sont plus adaptés à des tâches nécessitant une vaste connaissance linguistique, mais atteignent rarement les performances des modèles spécialisés. Nous observons que tous les modèles généralistes français obtiennent de meilleures performances uniquement

| Domaine | Modèle | DEFT-2021 | E3C | | Mantra-GSC | | | PxCorpus | QUAERO | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | *Temporal* | *Clinical* | *PATENTS* | *EMEA* | *MEDLINE* | | *EMEA* | *MEDLINE* |
| Général français | CamemBERT | 62,76** | 83,45 | 54,7** | 0,0** | 29,14** | 23,2** | 92,89** | 62,68** | 55,25** |
| | CamemBERTa | 62,61** | 83,22 | 55,53 | 44,16** | 40,84** | 22,55** | 95,05** | 64,86** | 55,6** |
| | FlauBERT | 33,51 | 61,64 | 47,61 | 31,47** | 66,2 | 20,69 | 47,57 | **74,86** | 48,98 |
| Biomédical français | DrBERT-7GB | 60,44** | 81,48** | 54,45 | **57,34** | 66,23 | 42,38 | 95,88* | 64,11** | 55,82** |
| | DrBERT-4GB | 61,52** | 83,14 | 55,14* | 49,93 | 62,88 | 41,84 | 95,91* | 65,58** | 56,01** |
| | DrBERT-CP | 63,43* | 83,43 | 56,55 | 39,68 | 60,88** | 35,52 | 71,38 | 67,05** | **60,1** |
| | CamemBERT-bio | **64,36** | 83,44 | **56,96** | 0,0** | 30,63** | 23,66** | 93,08** | 66,59** | 58,94 |
| Clinique français | ChuBERT-4GB | 59,77** | 82,97 | 54,32* | 47,66** | 67,75 | 33,23* | 95,71** | 58,26** | 51,55** |
| | ChuBERT-Mixed | 62,11** | **83,99** | 54,68** | 52,83 | **69,07** | **43,01** | **96,33** | 64,99** | 57,59** |
| Biomédical anglais | PubMedBERT | 60,27** | 80,86** | 38,34 | 4,51** | 40,14** | 27,53* | 94,66** | 53,19** | 53,26** |
| Général multilingue | XLM-RoBERTa | 60,32** | 82,6* | 52,87** | 8,58** | 52,64* | 18,73* | 95,8* | 64,47** | 51,12** |

TABLEAU 4.10 : Performances moyennes sur quatre exécutions pour les tâches de reconnaissance d'entités nommées. Le meilleur modèle est en gras et le second est souligné. La significativité statistique est calculée en utilisant le test de Student : * signifie $p < 0,05$, ** signifie $p < 0,01$.

| Domaine | Modèle | DEFT-2020 | DEFT-2021 | DiaMed | FrenchMedMCQA | MorFITT | PxCorpus |
|---|---|---|---|---|---|---|---|
| Général français | CamemBERT | 96,31 | 18,04** | 30,4** | 66,21 | 64,21** | 94,41 |
| | CamemBERTa | **97,96** | 18,04** | 24,05** | 64,44** | 66,28** | 93,95 |
| | FlauBERT | 42,37** | **39,21** | 34,08** | 61,88 | 70,25 | 93,45 |
| Biomédical français | DrBERT-7GB | 82,38 | 34,15** | 60,45 | 65,38 | 68,70** | 94,43 |
| | DrBERT-4GB | 74,42** | 33,99** | 57,56* | 65,75 | 68,19** | 94,08 |
| | DrBERT-CP | 95,71* | 30,04** | 54,43* | **66,22** | **70,99** | 94,52 |
| | CamemBERT-bio | 94,78* | 17,82** | 39,57** | 65,79 | 67,53** | 94,49 |
| Clinique français | ChuBERT-4GB | 50,87** | 38,79 | **62,82** | 64,26** | 65,15** | **95,12** |
| | ChuBERT-Mixed | 76,2* | 36,82** | 61,56 | 64,44** | 68,84** | 93,79 |
| Biomédical anglais | PubMedBERT | 95,33* | 25,53** | 54,96** | 65,41* | 68,58** | 93,12 |
| Général multilingue | XLM-RoBERTa | 67,66** | 24,46** | 26,69** | 64,69* | 67,28** | 93,91 |

TABLEAU 4.11 : Performances moyennes sur quatre exécutions pour les tâches de classification. Le meilleur modèle est en gras et le second est souligné. La significativité statistique est calculée en utilisant le test de Student : * signifie $p < 0,05$, ** signifie $p < 0,01$.

sur 3 des 20 tâches, mais restent néanmoins compétitifs sur la plupart des tâches. Les modèles pré-entraînés biomédicaux et cliniques montrent des performances supérieures dans les autres tâches. En particulier, `ChuBERT-Mixed` atteint la performance la plus élevée dans 4 tâches, suivi de `DrBERT-CP` et `CamemBERT-bio` avec 3 tâches, `ChuBERT-4GB` et `DrBERT-4GB` avec 2 tâches, puis `DrBERT-7GB` avec 1 tâche. Ces six modèles présentent les meilleures performances dans 75 % des tâches de DrBenchmark. Ces résultats soulignent l'importance de choisir un modèle adapté au domaine spécifique de la tâche pour obtenir de meilleures performances. En ce qui concerne la quantité de données utilisées pour le pré-entraînement des modèles (*small* vs. *large* ou *mixed*), les résultats montrent que plus les quantités de données sont grandes, plus les modèles sont performants, quelle que soit la stratégie de pré-entraînement ou la source de données (privée ou publique).

| Domaine | Modèle | SST | | MCQA |
|---|---|---|---|---|
| | | **CLISTER** | **DEFT-2020** | **FrenchMedMCQA** |
| Général français | CamemBERT | 0,55** / 0,33** | 0,58** / 0,51** | 28,53** / 2,25** |
| | CamemBERTa | 0,56** / 0,47** | 0,59** / 0,43** | 29,77** / 2,57** |
| | FlauBERT | 0,50** / 0,29** | 0,58** / 0,51** | 27,88** / 2,09** |
| Biomédical français | DrBERT-7GB | 0,62** / 0,57** | 0,72* / 0,81 | 31,07** / 3,22** |
| | DrBERT-4GB | 0,65* / 0,59** | 0,76 / 0,83 | 25,29** / 1,45** |
| | DrBERT-CP | 0,60* / 0,49* | 0,73 / **0,86** | 32,41** / 2,89** |
| | CamemBERT-bio | 0,54** / 0,26** | 0,58** / 0,32** | **35,30** / 1,45** |
| Clinique français | ChuBERT-4GB | 0,61** / 0,45** | 0,70** / 0,75* | 33,05** / **3,54** |
| | ChuBERT-Mixed | 0,63** / 0,53** | 0,68* / 0,80 | 26,08** / 0,96** |
| Biomédical anglais | PubMedBERT | **0,70** / **0,78** | **0,78** / **0,86** | 32,90** / 1,61** |
| Général multilingue | XLM-RoBERTa | 0,49** / 0,23** | 0,60** / 0,26* | 34,74** / 2,09** |

TABLEAU 4.12 : Performances moyennes sur quatre exécutions pour les tâches de similarité sémantique textuelle (SST) et de *question answering* (MCQA). Le meilleur modèle est en gras et le second est souligné. La significativité statistique est calculée en utilisant le test de Student : * signifie $p < 0,05$, ** signifie $p < 0,01$.

**Biomédical *vs.* Clinique.** À quantité de données égale, le modèle DrBERT-4GB, pré-entraîné sur le corpus NACHOS, obtient de meilleures performances dans 15 des 20 tâches, comparativement au modèle ChuBERT-4GB, qui a été pré-entraîné sur les données du CHU de Nantes. Ceci n'est pas surprenant, étant donné que la majorité des tâches de DrBenchmark proviennent du domaine biomédical. Le modèle ChuBERT-4GB excelle dans les tâches de classification de DEFT-2021 et DiaMed, qui sont composées de cas cliniques. En revanche, la combinaison des données biomédicales et cliniques pour le pré-entraînement d'un modèle semble favoriser un équilibre des performances. En effet, la comparaison entre DrBERT-7GB et ChuBERT-Mixed révèle que ce dernier surpasse le premier dans 13 des 20 tâches, comme illustré dans le tableau 4.13. Toutefois, les différences de performances entre les deux modèles sont minimes.

**Pré-entraînement de zéro *vs.* pré-entraînement poursuivi.** Les modèles DrBERT-CP et CamemBERT-bio, pré-entraînés respectivement à partir de PubMedBERT et de CamemBERT, montrent une amélioration des performances par rapport à leurs modèles initiaux. En se focalisant sur les modèles en pré-entraînement poursuivi, on observe que DrBERT-CP surpasse CamemBERT-bio dans 15 des 20 tâches. Ces résultats suggèrent que poursuivre le pré-entraînement d'un modèle anglais spécialisé dans le domaine cible (ici, PubMedBERT) semble être un meilleur choix qu'un modèle généraliste français (ici, CamemBERT). En comparant strictement DrBERT-7GB et DrBERT-CP, on observe que DrBERT-7GB est meilleur

que `DrBERT-CP` dans 9 des 20 tâches. De la même manière, on observe que `DrBERT-7GB` est meilleur que `CamemBERT-bio` sur 11 des 20 tâches. Ces observations indiquent qu'il n'y a finalement pas de réelles différences entre le pré-entraînement de zéro et le pré-entraînement poursuivi pour les modèles BERT.

Cependant, en tenant compte de l'impact environnemental, le pré-entraînement poursuivi semble être l'approche la plus responsable. Les émissions de carbone estimées pour le pré-entraînement de `CamemBERT-bio` s'élèvent à 0,84 kg d'équivalent CO2 (CO2eq), une mesure qui prend en compte les différents gaz à effet de serre et leur impact sur le réchauffement climatique. Il est important de noter que cette estimation n'inclut pas les émissions liées au modèle `CamemBERT`. En comparaison, le modèle `DrBERT-7GB` génère 36,9 kg de CO2eq, soit environ 44 fois plus. Puisque les modèles spécialisés sont développés à partir d'architectures existantes, il est plus écologique de poursuivre le pré-entraînement d'un modèle existant plutôt que d'en pré-entraîner un nouveau de zéro.

**Français *vs.* Autres langues.** Les modèles français obtiennent généralement de meilleures performances par rapport aux modèles anglais ou multilingues. Concernant le modèle anglais `PubMedBERT`, nous observons dans le tableau 4.13 que ses performances sont comparables à celles des modèles français dans la plupart des tâches, à l'exception des tâches de reconnaissance d'entités nommées où les modèles français sont supérieurs. Ainsi, nous constatons que la langue semble être moins prédominante lorsqu'elle est utilisée dans des tâches spécifiques à un domaine, comme celles du domaine biomédical.

**Architectures RoBERTa *vs.* DeBERTaV3.** Malgré un pré-entraînement sur seulement 30 % des données utilisées par $\texttt{CamemBERT}_{CCNET}$, `CamemBERTa` obtient des performances identiques ou supérieures dans 68 % des tâches (12 sur 20), bénéficiant de l'architecture DeBERTaV3 dans des scénarios spécifiques au domaine. Cependant, tous les modèles `CamemBERT` rencontrent des difficultés dans les corpus avec peu de données, comme MantraGSC Patents, où ils ne parviennent pas à générer d'étiquettes autres que 'O'. D'un autre côté, dans les mêmes scénarios à faibles ressources, les modèles `CamemBERTa` montrent une plus grande robustesse avec des performances plus élevées. L'architecture des modèles semble donc jouer un rôle dans les performances obtenues.

| Domaine | Modèle | POS *F1* | NER *F1* | CLS *F1* | SST *MSE / Spearman* | MCQA *Hamming / EMR* |
|---|---|---|---|---|---|---|
| Général français | CamemBERT | 96,57 | 51,56 | 61,60 | 0,57 / 0,42 | 28,53 / 2,25 |
| | CamemBERTa | 97,59 | 58,27 | 60,77 | 0,58 / 0,45 | 29,77 / 2,57 |
| | FlauBERT | 95,74 | 48,06 | 57,27 | 0,54 / 0,40 | 27,88 / 2,09 |
| Biomédical français | DrBERT-7GB | 97,97 | 64,24 | 67,70 | 0,67 / 0,69 | 31,07 / 3,22 |
| | DrBERT-4GB | **98,01** | 63,55 | 65,67 | 0,71 / 0,71 | 25,29 / 1,45 |
| | DrBERT-CP | 97,53 | 59,78 | **68,70** | 0,67 / 0,68 | 32,41 / 2,89 |
| | CamemBERT-bio | 96,71 | 53,07 | 63,34 | 0,56 / 0,29 | **35,30** / 1,45 |
| Clinique français | ChuBERT-4GB | 97,76 | 61,25 | 62,84 | 0,66 / 0,60 | 33,05 / **3,54** |
| | ChuBERT-Mixed | 97,95 | **64,96** | 66,94 | 0,66 / 0,67 | 26,08 / 0,96 |
| Biomédical anglais | PubMedBERT | 96,53 | 50,31 | 67,35 | **0,74** / **0,82** | 32,90 / 1,61 |
| Général multilingue | XLM-RoBERTa | 97,91 | 54,21 | 57,45 | 0,55 / 0,25 | 34,74 / 2,09 |

TABLEAU 4.13 : Performances moyennes par tâche pour les modèles BERT. Le meilleur modèle est en gras et le second est souligné.

#### 4.2.3.4 Analyse de la tokenisation des mots

Les *tokenizers* jouent un rôle crucial dans les modèles pré-entraînés, comme nous l'avons introduit dans la section 2.4.1. En raison des variations dans les données d'entraînement, les vocabulaires diffèrent selon les modèles, comme illustré dans la figure 4.1. En conséquence, les *tokenizers* segmentent les mots différemment, mais parviennent remarquablement à atteindre des niveaux de performance similaires comme précédemment noté dans le tableau 4.13.

Jusqu'à présent, il existait une idée répandue au sein de la communauté selon laquelle une segmentation excessive des mots avec les *tokenizers* entraînerait une perte de forme morphologique et de sémantique, introduisant ainsi du bruit et affectant négativement les performances des modèles (Church, 2020; Bostrom and Durrett, 2020; Hofmann et al., 2021). Cependant, comme le montre le tableau 4.14, nos expériences révèlent que `FlauBERT` est le modèle qui segmente le moins les mots (1,45 *tokens* par mot en moyenne), tandis que `DrBERT-CP` a tendance à plus segmenter les mots (1,96 *tokens* par mot en moyenne). Même si `DrBERT-CP` est un modèle biomédical et `FlauBERT` un modèle général, en comparant les performances de ces deux modèles sur les tâches de DrBenchmark, nous observons que `DrBERT-CP` surpasse `FlauBERT` sur 16 des 20 tâches, contredisant ainsi les conclusions émises par la communauté sur la performance de ces modèles.

Le tableau 4.15 donne quelques exemples de segmentation en tokens d'une liste de termes biomédicaux et cliniques couramment utilisés. Les termes sont peu découpés par le modèle `ChuBERT-4GB` ce qui montre que ces termes sont très présents dans le corpus de pré-entraînement. Malgré les éléments d'analyse rapportés, il est difficile de tirer des

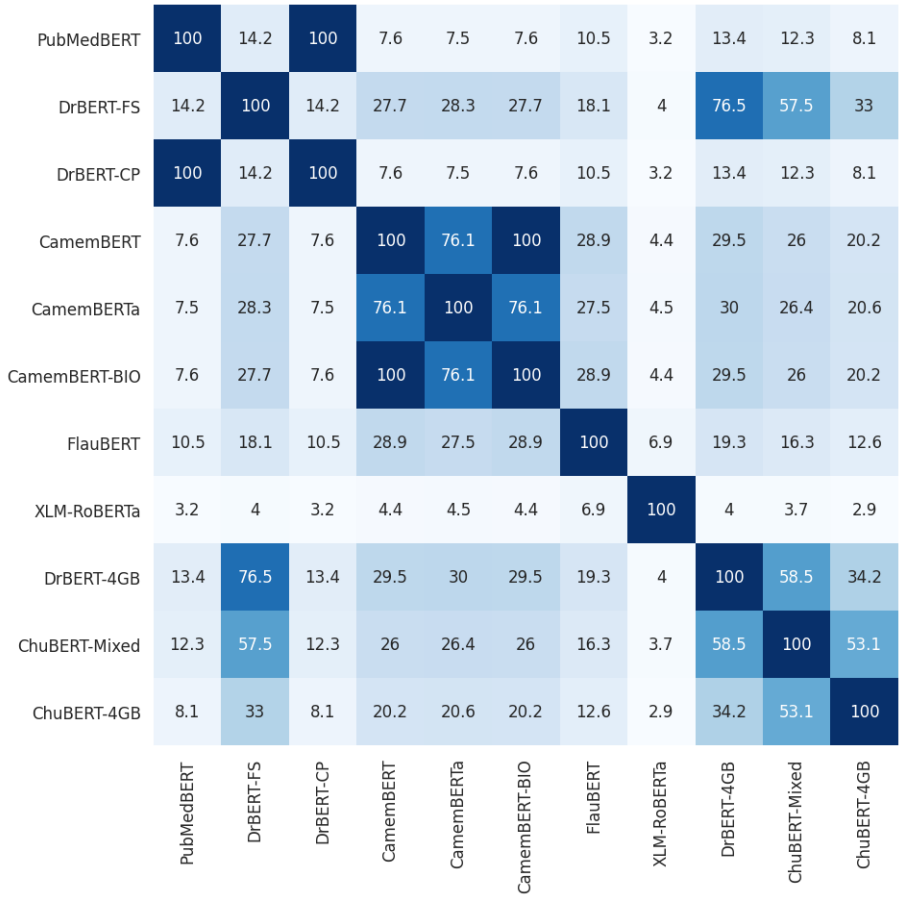


FIGURE 4.1 : Matrice d'intersection des vocabulaires.

conclusions définitives sur l'impact des *tokenizers* dans les modèles. En effet, de nombreux paramètres peuvent amener à des variations de performances : algorithmes de *tokenization* utilisé, nature des données, quantité, etc. Ainsi, il est nécessaire d'examiner spécifiquement l'impact de ces algorithmes sur la construction des modèles avec une étude plus approfondie dédiée à ce sujet.

| | Général français | | | Biomédical français | | | Clinique français | | Biomédical anglais | Général multilingue |
|---|---|---|---|---|---|---|---|---|---|---|
| **Corpus - Tâche** | **CamemBERT** | **CamemBERTa** | **FlauBERT** | **DrBERT-7GB** | **DrBERT-CP** | **CamemBERT-bio** | **ChuBERT-4GB** | **ChuBERT-Mixed** | **PubMedBERT** | **XLM-RoBERTa** |
| CAS - POS | 1,63 | 1,64 | 1,34 | 1,36 | 1,81 | 1,63 | 1,33 | **1,32** | 1,81 | 1,8 |
| ESSAI - POS | 1,55 | 1,56 | **1,28** | 1,29 | 1,78 | 1,55 | 1,38 | 1,3 | 1,78 | 1,75 |
| QUAERO - NER EMEA | 1,66 | 1,67 | **1,37** | **1,37** | 1,73 | 1,66 | 1,46 | 1,39 | 1,73 | 1,77 |
| QUAERO - NER Medline | 2,01 | 2,01 | **1,58** | 1,64 | 1,97 | 2,01 | 1,69 | 1,63 | 1,97 | 2,18 |
| E3C - NER FR Clinical | 1,64 | 1,65 | 1,39 | 1,32 | 1,8 | 1,64 | **1,28** | **1,28** | 1,8 | 1,78 |
| E3C - NER FR Temporal | 1,64 | 1,65 | 1,39 | 1,32 | 1,81 | 1,64 | **1,28** | **1,28** | 1,81 | 1,79 |
| MORFITT - CLS | 1,51 | 1,51 | **1,33** | 1,39 | 1,91 | 1,51 | 1,52 | 1,4 | 1,91 | 1,73 |
| FrenchMedMCQA - MCQA | 1,8 | 1,8 | **1,55** | 1,55 | 2,03 | 1,8 | 1,67 | 1,57 | 2,03 | 2,0 |
| FrenchMedMCQA - CLS | 1,8 | 1,8 | **1,55** | 1,55 | 2,03 | 1,8 | 1,67 | 1,57 | 2,03 | 2,0 |
| MANTRAGSC - NER FR EMEA | 1,5 | 1,46 | **1,34** | 1,37 | 1,99 | 1,5 | 1,41 | 1,38 | 1,99 | 1,71 |
| MANTRAGSC - NER FR Medline | 2,25 | 2,25 | **1,88** | 2,05 | 2,47 | 2,25 | 2,15 | 2,04 | 2,47 | 2,49 |
| MANTRAGSC - NER FR Patents | 1,58 | 1,58 | **1,41** | 1,51 | 2,06 | 1,58 | 1,67 | 1,52 | 2,06 | 1,86 |
| CLISTER - SST | 1,76 | 1,76 | 1,55 | 1,55 | 2,09 | 1,76 | **1,49** | **1,49** | 2,09 | 1,93 |
| DEFT-2020 - SST | 1,43 | 1,43 | **1,31** | 1,45 | 1,92 | 1,43 | 1,57 | 1,48 | 1,92 | 1,64 |
| DEFT-2020 - CLS | 1,31 | 1,32 | **1,2** | 1,23 | 1,75 | 1,31 | 1,29 | 1,22 | 1,75 | 1,51 |
| DEFT-2021 - CLS | 1,7 | 1,71 | 1,48 | 1,51 | 2,05 | 1,7 | **1,45** | **1,45** | 2,05 | 1,9 |
| DEFT-2021 - NER | 1,62 | 1,63 | 1,35 | 1,35 | 1,8 | 1,62 | **1,31** | **1,31** | 1,8 | 1,79 |
| DiaMed - CLS | 1,66 | 1,67 | 1,45 | 1,46 | 1,99 | 1,66 | **1,38** | **1,38** | 1,99 | 1,88 |
| PxCorpus - NER | 1,71 | 1,76 | 1,63 | 1,66 | 2,13 | 1,71 | **1,61** | 1,68 | 2,13 | 1,83 |
| PxCorpus - CLS | 1,71 | 1,76 | 1,63 | 1,66 | 2,13 | 1,71 | **1,61** | 1,68 | 2,13 | 1,83 |
| Moyenne | 1,67 | 1,68 | **1,45** | 1,48 | 1,96 | 1,67 | 1,51 | 1,47 | 1,96 | 1,86 |

TABLEAU 4.14 : Nombre moyen de tokens par mot pour chaque modèle et corpus. Pour chaque tâche, le nombre de tokens moyen le plus bas est affiché en gras, et le nombre le plus élevé est souligné. Les cellules en vert indiquent le meilleur modèle en termes de performance pour la tâche, tandis que les cellules en rouge indiquent le modèle le moins performant.

| | Général français | | Biomédical français | Clinique français | | Biomédical anglais | Général multilingue |
|---|---|---|---|---|---|---|---|
| **Terme** | **CamemBERT CamemBERT-BIO** | **FlauBERT** | **DrBERT-7GB** | **ChuBERT-4GB** | **ChuBERT-Mixed** | **PubMedBERT DrBERT-CP** | **XLM-RoBERTa** |
| *asymptomatique* | a-s-y-mp-to-matique | as-ym-ptom-atique | ✓ | ✓ | ✓ | asympt-omat-ique | as-y-mp-tomat-ique |
| *blépharorraphie* | blé-phar-or-ra-phi-e | bl-é-phar-or-raph-ie | blé-ph-ar-or-ra-ph-ie | bléphar-or-raphie | blé-ph-ar-or-ra-phie | ble-pha-ror-ra-phi-e | b-lép-har-orra-phi-e |
| *bradycardie* | brad-y-cardi-e | bra-dy-car-die | ✓ | ✓ | ✓ | brady-car-di-e | bra-dy-card-ie |
| *bronchographie* | bron-ch-ographie | bron-cho-graphie | bronch-ographie | bron-chographie | bronch-ographie | bronch-ograph-ie | bron-ch-ographie |
| *bronchopneumopathie* | bron-cho-p-ne-um-opathie | bron-chop-neu-mo-pathie | bronchop-neumopathie | bronch-opneumopathie | bronch-opneumopathie | bronch-op-neum-opath-ie | bron-chop-ne-umo-pathi-e |
| *dysménorrhée* | dys-mén-or-r-h-ée | dys-mé-nor-rh-ée | dys-m-énorrhée | ✓ | dysm-énorrhée | dysm-eno-rr-he-e | dys-mén-or-r-hé-e |
| *glaucome* | gla-uc-ome | glau-come | ✓ | ✓ | ✓ | glauc-ome | gla-u-come |
| *IRM* | ✓ | ✓ | ✓ | ✓ | ✓ | ir-m | I-RM |
| *kystectomie* | ky-st-ectomie | ky-st-ec-tomie | kys-tectomie | ✓ | kyste-ctomie | ky-st-ectom-ie | ky-st-ecto-mie |
| *neuroleptique* | neuro-le-p-tique | neur-ol-ep-tique | neur-oleptique | ✓ | neurole-ptique | neurol-ept-ique | neuro-lep-tique |
| *nicotine* | ✓ | ✓ | ✓ | nic-otine | ✓ | ✓ | nico-tine |
| *poliomyélite* | poli-om-y-élite | poli-omy-élite | poli-omyélite | pol-i-omyélite | poli-omyélite | poli-omyel-ite | poli-om-y-é-lite |
| *rhinopharyngite* | rhin-oph-ary-ng-ite | rh-ino-phar-yn-gite | rhin-opharyng-ite | ✓ | ✓ | rhin-oph-aryng-ite | r-hin-op-har-y-ng-ite |
| *toxicomanie* | toxico-mani-e | ✓ | ✓ | toxic-omanie | toxic-omanie | toxic-oman-ie | toxic-om-anie |
| *vasoconstricteur* | vas-oc-on-strict-eur | vas-o-cons-tri-cteur | vasoconstric-teur | vas-ocons-tric-teur | vasoconstric-teur | vasoconstric-te-ur | vaso-con-strict-eur |

TABLEAU 4.15 : Comparaison des tokenizers des modèles sur des termes biomédicaux couramment utilisés. Le symbole ✓ indique que le mot est présent en tant que token entier dans le vocabulaire, tandis que des tirets séparent les tokens composant un mot.

## 4.3 Adaptation de modèles Longformer au domaine biomédical français pour les longs documents

Les documents cliniques sont généralement des documents de grande taille pouvant contenir jusqu'à plusieurs milliers de mots. Les modèles BERT, limité à 512 tokens en entrée, ne peuvent pas être appliqués sur les documents entiers, à moins de segmenter le document et traiter chaque segment séparément. Cependant, avec cette approche, il n'y a pas de mécanisme d'auto-attention entre les segments traités en parallèle, ce qui peut avoir un impact sur certaines tâches où de longues dépendances sont présentes au sein du document. Plusieurs travaux ont démontré l'intérêt des modèles Longformer pour représenter

les longs documents cliniques anglais (Li et al., 2023). Dans cette section, nous proposons plusieurs stratégies d'adaptation de modèles Longformer pour le domaine médical français. Plus précisément, nous introduisons les modèles `DrLongformer-FS` (pré-entraîné de zéro sur le corpus NACHOS), `DrLongformer-CP` (pré-entraînement poursuivi du modèle `Clinical-Longformer` sur le corpus NACHOS) et `DrBERT-4096` (conversion du modèle `DrBERT-7GB` en Longformer). L'évaluation de ces trois nouveaux modèles a été réalisée sur un ensemble de tâches constituées de longues séquences issues de DrBenchmark. Les performances des modèles Longformer ont été comparées aux modèles BERT afin de vérifier si un plus grand contexte améliore les performances.

### 4.3.1 Données de pré-entraînement

Nous avons réutilisé le corpus NACHOS dans sa version document intégral. Le corpus de pré-entraînement est composé de à 2 367 419 documents pour un total d'environ 1 276M de mots. Les tokenizers de chaque modèle ont été appliqués au corpus pour segmenter les documents en tokens. Chaque document du corpus a ensuite été découpé en multiple de 4 096 tokens. Les séquences inférieures à 512 tokens en fin de documents n'ont pas été conservées et celles supérieures à 512 tokens étaient complétées avec du *padding*.

Pour le modèle pré-entraîné de zéro `DrLongformer-FS`, nous entraînons un tokenizer WordPiece sur le corpus de pré-entraînement avec une taille de vocabulaire de 50 265 tokens. Pour les modèles `DrLongformer-CP` et `DrBERT-4096`, le tokenizer utilisé est identique aux modèles `Clinical-Longformer` et `DrBERT-7GB` respectivement.

### 4.3.2 Pré-entraînement des modèles

Comme nous l'avons fait pour les travaux sur `DrBERT`, nous comparons la stratégie de pré-entraînement afin de vérifier si les résultats observés avec les modèles Longformer corroborent ceux observés avec les modèles BERT. Nous proposons ici d'évaluer trois stratégies de pré-entraînement :

- Pré-entraîner un modèle de zéro, incluant le tokenizer à partir des données de pré-entraînement
- Poursuivre le pré-entraînement du modèle anglais clinique `Clinical-Longformer` sur les données en français en conservant le tokenizer initial.

- Convertir le modèle `DrBERT-7GB` en Longformer en remplaçant les poids du Longformer avec ceux de `DrBERT-7GB`.

Comme il n'existe pas de modèle `Longformer` général pour le français, nous pouvons utiliser seulement les modèles anglais comme point de départ pour le pré-entraînement poursuivi.

| Modèle | Architecture | Stratégie de pré-entraînement | Corpus |
|---|---|---|---|
| `DrLongformer-FS` | Longformer | À partir de zéro | NACHOS |
| `DrLongformer-CP` | Longformer | Pré-entraînement poursuivi (`Clinical-Longformer`) | NACHOS |
| `DrBERT-4096` | Longformer | Conversion (`DrBERT-4096`) | NACHOS |

Tableau 4.16 : Liste des configurations des modèles pré-entraînés.

**Optimisation et pré-entraînement** Tous les modèles pré-entraînés reprennent l'architecture Longformer (12 couches, 768 dimensions cachées, 12 têtes d'attention, 149M de paramètres). Les modèles sont pré-entraînés sur la tâche de *Masked Language Modeling* (MLM) avec une probabilité de masquage de 15 %. Les modèles sont optimisés pendant environ 50k étapes (*steps*) avec une taille de lot (*batch size*) de 256 séquences, chaque séquence étant remplie de 4 096 tokens permettant de traiter environ un million de *tokens* par étape. À l'instar du pré-entraînement des modèles `DrBERT`, le taux d'apprentissage (*learning rate*) est progressivement augmenté de manière linéaire pendant 10k étapes, passant de zéro au taux d'apprentissage initial de $5\times10^{-5}$. Les modèles sont entraînés sur 128 GPUs Nvidia V100 de 32 GB pendant 20 heures sur le supercalculateur Jean Zay. Nous utilisons également l'entraînement à précision mixte (FP16) (Micikevicius et al., 2017) pour réduire l'empreinte mémoire.

Après avoir adapté le modèle `DrBERT-7GB` à l'architecture Longformer, aboutissant au nouveau modèle nommé `DrBERT-4096`, nous l'avons pré-entrainé pour une époque supplémentaire sur NACHOS afin d'adapter les poids à des séquences plus longues. Nous avons effectué 13 300 étapes sur 22 heures en utilisant une seule GPU Tesla A100 avec une taille de lot de 8 et une accumulation de gradient sur 4 étapes, résultant en une taille de lot effective de 32.

### 4.3.3 Évaluation des modèles

L'évaluation des modèles Longformer a été réalisée sur une partie des tâches de DrBenchmark sélectionnées sur deux critères. Premièrement, nous avons sélectionné les tâches composées de documents longs, c'est-à-dire ayant des séquences supérieures à 250 mots. En effet, en considérant une segmentation moyenne d'un mot à deux tokens comme indiqué dans le tableau 4.14, une séquence de 250 mots ne pourra pas être contenue intégralement dans un modèle `BERT` et un modèle avec un plus grand contexte est donc nécessaire. Sur ce critère, nous avons sélectionné les corpus DEFT-2021, DiaMed, E3C, QUAERO-EMEA, MorFITT et la tâche de classification du corpus DEFT-2020. Tous ces corpus portent sur les tâches de classification de textes et de reconnaissance d'entités nommées. Afin d'ajouter de la diversité dans les tâches d'évaluation, nous avons ensuite ajouté les corpus CAS, ESSAI, FrenchMedMCQA, CLISTER et la tâche de similarité de textes du corpus DEFT-2020. L'ensemble des corpus d'évaluation sélectionnés sont présentés dans le tableau 4.17. Pour les tâches de reconnaissance d'entités nommées des corpus DEFT-2021, QUAERO et E3C, l'évaluation porte sur les documents entiers et non sur les phrases après segmentation des documents comme présenté dans DrBenchmark.

| Corpus | Tâche | Métrique | Train | Dev. | Test | Long. Moy. | Long. Max. |
|---|---|---|---|---|---|---|---|
| ESSAI | POS Tagging | SeqEval F1 | 5 072 | 725 | 1 450 | 22,55 | 135 |
| CAS Corpus | POS Tagging | SeqEval F1 | 2 653 | 379 | 758 | 23,07 | 136 |
| CLISTER | Similarité textes | MSE/Spearman | 499 | 101 | 400 | 25,21 | 103 |
| DEFT-2020 | Similarité textes | MSE/Spearman | 498 | 102 | 410 | 45,14 | 118 |
| DEFT-2020 | Classification | Weighted F1 | 460 | 112 | 530 | 90,19 | 254 |
| DEFT-2021 | Classification | Weighted F1 | 118 | 49 | 108 | 404,78 | 1 710 |
| DEFT-2021 | REN | SeqEval F1 | 121 | 46 | 108 | 404,78 | 1 710 |
| QUAERO - EMEA | REN | SeqEval F1 | 13 | 12 | 15 | 1 006,42 | 1 549 |
| E3C - Clinical | REN | SeqEval F1 | 146 | 22 | 81 | 356,39 | 714 |
| E3C - Temporal | REN | SeqEval F1 | 56 | 8 | 17 | 354,86 | 685 |
| FrenchMedMCQA | MCQA | EMR/Hamming Score | 2 171 | 312 | 622 | 44,80 | 174 |
| FrenchMedMCQA | Classification | Weighted F1 | 2 171 | 312 | 622 | 44,80 | 174 |
| MorFITT | Classification | Weighted F1 | 1 514 | 1 022 | 1 088 | 226,33 | 1 425 |
| DiaMed | Classification | Weighted F1 | 509 | 76 | 154 | 329,92 | 1 379 |

TABLEAU 4.17 : Corpus, tâches et métriques utilisés pour évaluer les modèles pour les longs documents. **Long. Moy.** correspond à la longueur moyenne des séquences et **Long. Max.** à la longueur maximum des séquences.

**Modèles de références** Les trois modèles Longformer introduits dans ce travail sont comparés à cinq autres modèles : `DrBERT-7GB`, `CamemBERT-bio`, `ChuBERT-4GB`, `ChuBERT-Mixed` et `Clinical-Longformer`. Les performances des modèles BERT sont extraites de la section 4.2, tandis que celles du modèle `Clinical-Longformer` sont obtenues après

une phase d'affinage réalisée dans la même configuration que pour les autres modèles Longformer.

**Affinage et évaluation** Tous les modèles Longformer sont affinés de la même manière à partir des pipelines HuggingFace propres à chaque tâche. Pour les tâches de reconnaissance d'entités nommées et d'étiquetage morphosyntaxique, les modèles BERT sont affinés au niveau phrase tandis que les modèles Longformer sont affinés au niveau document (ou la moitié du document lorsqu'il dépasse 4 096 tokens). Pour les tâches de classification de textes, les documents sont tronqués aux 4 096 premiers tokens pour correspondre à la limite de longueur des modèles Longformer. Pour les modèles BERT, nous tronquons les documents aux 512 premiers tokens.

Plusieurs taux d'apprentissage (*learning rate*) sont expérimentés lors de l'affinage des modèles : $\{1e-4, 1e-5, 2e-5, 5e-5\}$. Quatre exécutions sont réalisées pour chaque taux d'apprentissage et nous conservons les résultats du taux d'apprentissage qui obtient la meilleure moyenne sur les quatre exécutions.

### 4.3.4 Résultats

Les résultats de chaque modèle sur les tâches sont présentés par type de tâche dans les tableaux 4.18 pour les tâches de reconnaissance d'entités nommées, 4.19 pour les tâches de classification et 4.20 pour les tâches d'étiquetage morphosyntaxique, de similarité sémantique textuelle et de *question answering*. Nous proposons d'analyser ces résultats sous deux angles : (i) les stratégies de pré-entraînement et (ii) l'architecture du modèle, en considérant à la fois la longueur maximale de la séquence d'entrée et le mécanisme d'attention.

| Modèle | DEFT-2021 | E3C - Clinical | E3C - Temporal | QUAERO - EMEA |
|---|---|---|---|---|
| DrBERT-7GB | 60,44* | 54,45 | 81,48* | 64,11 |
| CamemBERT-bio | **64,36** | 56,96 | <u>83,44</u> | **66,59** |
| ChuBERT-4GB | 59,77* | 54,32* | 82,97 | 58,26* |
| ChuBERT-Mixed | 62,11* | 54,68* | **83,99** | 64,99 |
| Clinical-Longformer | 55,62* | 50,88* | 79,11* | 50,36* |
| DrBERT-4096 | 59,74* | **57,20** | 81,98* | 64,09 |
| DrLongformer-CP | <u>63,73*</u> | <u>57,07</u> | 83,32 | 64,78 |
| DrLongformer-FS | 55,00* | 55,74 | 79,23* | <u>66,31</u> |

TABLEAU 4.18 : Performances des modèles sur les tâches de reconnaissance d'entités nommées. Le meilleur modèle est en gras et le second est souligné. La significativité statistique est calculée en utilisant le test de Student : * signifie $p < 0{,}05$.

| Modèle | DEFT-2020 | DEFT-2021 | DiaMed | FrenchMedMCQA | MorFITT |
|---|---|---|---|---|---|
| DrBERT-7GB | 82,38 | 34,15* | 60,45 | 65,38 | 68,70* |
| CamemBERT-bio | 94,78 | 17,82* | 39,57* | 65,79 | 67,53* |
| ChuBERT-4GB | 50,87* | 38,79* | **62,82** | 64,26* | 65,15* |
| ChuBERT-Mixed | 76,2* | 36,82* | 61,56 | 64,44* | 68,84* |
| Clinical-Longformer | **95,59** | 27,73* | 39,91* | 64,57 | 62,98* |
| DrBERT-4096 | 87,48 | 36,98* | 61,17 | **66,59** | 70,75* |
| DrLongformer-CP | 94,27 | **51,23** | 54,82* | 66,00 | **71,95** |
| DrLongformer-FS | 36,57* | 49,54 | 54,05* | 64,42 | 69,94* |

TABLEAU 4.19 : Performances des modèles sur les tâches de classification. Le meilleur modèle est en gras et le second est souligné. La significativité statistique est calculée en utilisant le test de Student : * signifie $p < 0,05$.

| | POS | | SST | | MCQA |
|---|---|---|---|---|---|
| Modèle | CAS | ESSAI | CLISTER | DEFT-2020 | FrenchMedMCQA |
| DrBERT-7GB | 96,93* | 98,41* | 0,62 / 0,57* | 0,72 / 0,81* | 31,07* / 3,22* |
| CamemBERT-bio | 95,22* | 97,39* | 0,54* / 0,26* | 0,58* / 0,32* | **35,30** / 1,45* |
| ChuBERT-4GB | 97,05* | 98,46* | 0,61 / 0,45* | 0,63* / 0,53* | 33,05* / **3,54** |
| ChuBERT-Mixed | 97,30* | 98,60* | 0,63 / 0,53* | 0,68* / 0,80* | 26,08* / 0,96* |
| Clinical-Longformer | 97,26* | 98,54* | 0,60 / **0,88** | 0,71 / 0,85 | 28,55* / 1,45* |
| DrBERT-4096 | 97,55* | 98,60* | **0,65** / 0,71* | 0,73 / 0,84 | 28,20* / **3,54** |
| DrLongformer-CP | **97,73** | **98,72** | 0,60 / 0,86 | **0,78** / **0,88** | 28,62* / 1,45* |
| DrLongformer-FS | 97,29* | 98,48* | 0,64 / 0,58* | 0,65 / 0,79* | 23,50* / 3,38* |

TABLEAU 4.20 : Performances des modèles sur les tâches d'étiquetage morphosyntaxique (POS), similarité sémantique textuelle (SST) et de *question answering* (MCQA). Le meilleur modèle est en gras et le second est souligné. La significativité statistique est calculée en utilisant le test de Student : * signifie $p < 0,05$.

#### 4.3.4.1 Impact des stratégies de pré-entraînement

Contrairement à ce que nous avons observé avec les modèles BERT, l'utilisation d'un modèle existant pour le pré-entraînement — que ce soit par la conversion d'un modèle BERT ou la poursuite du pré-entraînement d'un modèle Longformer — se révèle donner de meilleurs résultats que de pré-entraîner un modèle Longformer de zéro. En effet, le modèle `DrLongformer-FS` ne surpasse jamais les performances de `DrBERT-4096` ou `DrLongformer-CP`. Les résultats révèlent également que `DrLongformer-FS` est supérieur à `Clinical-Longformer` dans seulement 10 des 16 tâches évaluées. Toutefois, il est crucial de souligner qu'aucun des modèles Longformer mentionnés dans la littérature n'a été pré-entraîné de zéro. Par exemple, `Clinical-Longformer` a bénéficié d'un pré-entraînement à

| Modèle | POS *F1* | CLS *F1* | NER *F1* | SST *MSE / Spearman* | MCQA *Hamming / EMR* |
|---|---|---|---|---|---|
| DrBERT-7GB | 97,67 | 62,21 | 65,12 | 0,67 / 0,69 | 31,07 / 3,22 |
| CamemBERT-bio | 96,31 | 57,10 | **67,84** | 0,56 / 0,29 | **35,30** / 1,45 |
| ChuBERT-4GB | 97,76 | 56,38 | 63,83 | 0,62 / 0,49 | 33,05 / **3,54** |
| ChuBERT-Mixed | 97,95 | 61,57 | 66,44 | 0,66 / 0,67 | 26,08 / 0,96 |
| Clinical-Longformer | 97,90 | 58,16 | 58,99 | 0,66 / **0,87** | 28,55 / 1,45 |
| DrBERT-4096 | 98,08 | 64,60 | 65,75 | **0,69** / 0,78 | 28,20 / **3,54** |
| DrLongformer-CP | **98,23** | **67,65** | 67,22 | **0,69** / **0,87** | 28,62 / 1,45 |
| DrLongformer-FS | 97,89 | 54,90 | 64,07 | 0,65 / 0,69 | 23,50 / 3,38 |

TABLEAU 4.21 : Performances moyennes par tâche pour les modèles Longformer. Le meilleur modèle est en gras et le second est souligné.

partir des poids de Longformer, qui lui-même a été pré-entraîné à partir de RoBERTa. En définitive, le modèle DrLongformer-CP a bénéficié de quatre phases de pré-entraînement, acquérant ainsi une robustesse accrue grâce à la diversité des données vues au cours de ces multiples phases. Pour une évaluation équitable, DrLongformer-FS aurait besoin de passer par une durée de pré-entraînement équivalente à celle des modèles ayant bénéficié de plusieurs phases de pré-entraînement. Cependant, une telle approche s'avérerait excessivement coûteuse, sans offrir de garantie d'amélioration significative des performances.

Dans la comparaison entre DrBERT-4096 et DrLongformer-CP, les résultats sont plus nuancés. DrLongformer-CP, pré-entraîné à partir des poids de Clinical-Longformer, affiche de meilleures performances dans 11 des 16 tâches. Le modèle DrLongformer-CP semble tirer avantage des données cliniques issues de la base de données MIMIC, utilisées pour le pré-entraînement de Clinical-Longformer. Cette observation corrobore nos conclusions sur les modèles BERT concernant l'efficacité du transfert de connaissances spécialisées entre les langues, particulièrement lorsqu'il s'agit de poursuivre le pré-entraînement d'un modèle clinique anglais avec des données biomédicales en français. Cependant, nous ne savons pas si cette approche est plus avantageuse que poursuivre le pré-entraînement d'un modèle français puisqu'il n'existe aucun modèle Longformer français pour réaliser cette expérience.

#### 4.3.4.2 Impact de l'architecture sur les tâches d'évaluation

En ce qui concerne les architectures des modèles pré-entraînés, les résultats montrent que les modèles Longformer sont plus performants que les modèles BERT sur 8 des 14

tâches. En analysant les résultats en fonction du type de tâche d'évaluation à partir des tableaux 4.19, 4.18 et 4.20, nous pouvons tirer différentes conclusions.

**Reconnaissance d'entités nommées** Nous remarquons que les modèles BERT sont plus adaptés aux tâches de reconnaissance d'entités nommées (REN). En effet, bien que les modèles BERT traitent les tâches de REN phrase à phrase et non le document entier, ils obtiennent de meilleurs résultats que les modèles Longformer sur 3 des 4 tâches de REN. Cela montre qu'il n'est pas nécessaire d'avoir une grande taille de contexte pour cette tâche et que le contexte de la phrase courante est généralement suffisant dans les domaines biomédical et clinique. Bien que les modèles BERT affichent des performances supérieures sur les tâches de REN individuellement, le tableau 4.21 montre que `DrLongformer-CP` obtient un score F1 moyen plus élevé sur l'ensemble des tâches, démontrant ainsi la robustesse du modèle.

**Classification de textes** Les tâches de classification de textes semblent tirer profit de la taille de contexte de 4 096 tokens des Longformer. Pour les tâches composées de séquences supérieures à 512 tokens, comme DEFT-2021 et MorFITT, les résultats montrent une nette avantage des modèles Longformer avec en moyenne un gain de 7,78 en score F1 entre le meilleur modèle BERT et le meilleur modèle Longformer. Pour les autres tâches de classification composées de séquences plus courtes (FrenchMedMCQA et DEFT-2020), la même tendance est observée, bien que les écarts de performances entre les modèles des deux architectures soient moins prononcés, avec un gain moyen de score F1 de seulement 0,81. Sur l'ensemble des tâches, on note un gain en score F1 de 5,44 entre `DrBERT-7GB` et `DrLongformer-CP`.

En analysant la distribution de l'attention sur les documents du jeu de test de DEFT-2021 dans la figure 4.2, nous observons que l'attention du modèle est répartie uniformément sur l'ensemble du document. Les taux d'erreur dans le tableau 4.22 montrent que les modèles Longformer surpassent DrBERT tant sur les documents courts que les longs.

**Similarité sémantique textuelle** Pour les tâches de similarité sémantique textuelle (DEFT-2020 et CLISTER), les résultats rapportés sont très proches entre les modèles et aucune architecture ne se démarque entre BERT et Longformer. En termes de corrélation de Spearman, il n'y a pas de différence significative entre BERT et Longformer.

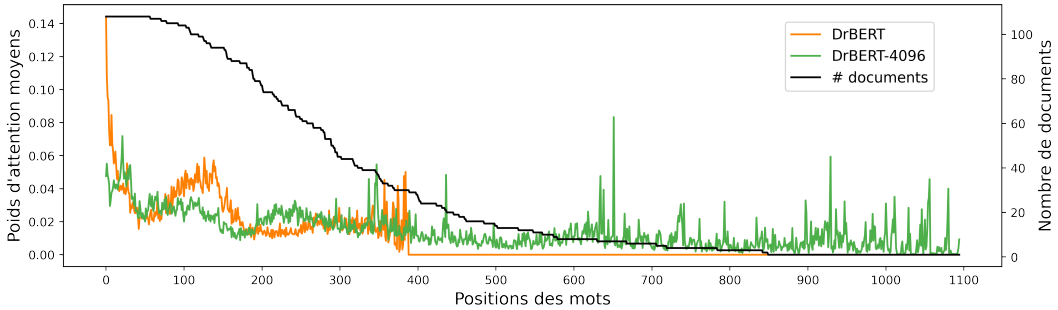


FIGURE 4.2 : Poids moyens d'attention pour chaque position de mot des documents du jeu de test du corpus DEFT-2021. Les poids d'attention sont obtenus à partir du token `[CLS]` utilisé pour la classification en additionnant les poids d'attention de toutes les têtes d'attention de la dernière couche du modèle. Ces poids finaux reflètent l'importance relative de chaque mot après avoir intégré des informations contextuelles provenant des autres mots dans la phrase au travers des 12 couches d'auto-attention du modèle transformer. À noter que ces poids d'attention sont une représentation moyenne, calculée sur l'ensemble des instances de l'ensemble de test, fournissant ainsi une vue d'ensemble de l'importance attribuée à chaque position de mot en moyenne.

| | **DEFT-2021** | |
|---|---|---|
| Taille du document en mots | $\leq 388$ | $> 388$ |
| Nb. documents dans le test set | 78 | 30 |
| `DrBERT-7GB` | 13,38 | 21,01 |
| `DrBERT-4096` | 12,82 | 20,58 |
| `DrLongformer-CP` | 11,32 | 16,96 |

TABLEAU 4.22 : Taux d'erreurs de `DrBERT-7GB`, `DrBERT-4096` et `DrLongformer-CP` sur le jeu de test du corpus DEFT-2021. Le seuil de longueur de séquence de mots sélectionné correspond à la limite maximale de longueur pour `DrBERT-7GB`.

**Questions à choix multiples** Enfin, sur la tâche de questions à choix multiples du corpus FrenchMedMCQA, les modèles BERT semblent plus efficaces avec des scores d'EMR supérieurs à 30 % contre 28,62 % maximum pour le meilleur des modèles Longformer.

## 4.4 Empreinte carbone des modèles de langue pré-entraînés

Depuis l'avènement des techniques d'apprentissage profond et le succès de l'architecture Transformer, les modèles de langue pré-entraînés ont connu une amélioration drastique en matière de tâches de compréhension du langage, surpassant parfois les performances humaines. Au fil des années, la disponibilité de données à grande échelle a permis le développement de nouveaux modèles sur différentes architectures. Depuis la parution de ChatGPT en novembre 2022, nous avons assisté à une prolifération de grands modèles de langue toujours plus volumineux que les précédents, atteignant jusqu'à plusieurs dizaines de milliards de paramètres. Bien que ces modèles soient au cœur des discussions chez les chercheurs et les professionnels de l'industrie, leur impact environnemental est rarement discuté. En effet, cette croissance soudaine du nombre de modèles pré-entraînés s'accompagne inévitablement d'une croissance de l'empreinte carbone générée par ces modèles. Même si les progrès des logiciels et du matériel ont considérablement réduit les coûts de pré-entraînement des modèles, leur empreinte carbone sur l'environnement reste non négligeable. L'impact écologique de ces modèles se retrouve à plusieurs niveaux : (i) lors du pré-entraînement, (ii) à l'utilisation, lors de l'inférence, et plus globalement (iii) tout au long du cycle de vie (réseau, stockage, etc.) (Khowaja et al., 2023). Pour les grands modèles de langue basés sur Transformer, on estime à 6 FLOPS (le nombre d'opérations en virgule flottante par seconde) par paramètre le coût pour s'entraîner sur un token contre 1 à 2 FLOPS par paramètre pour inférer sur un token (Kaplan et al., 2020).

Les travaux présentés dans ce chapitre portent sur des modèles avec un nombre de paramètres moins important. Pour autant, le pré-entraînement de tels modèles nécessite une grande quantité de ressources de calcul. La majorité des travaux présentés précédemment ont été réalisés sur le supercalculateur Jean Zay ou le Centre de calcul intensif des Pays de la Loire (CCIPL).

Pour les travaux portant sur DrBERT, environ 18 000 heures de calcul GPU ont été nécessaires pour créer les 7 modèles présentés dans l'étude, environ 7 500 heures de calcul GPU pour le débogage en raison de problèmes techniques liés aux configurations des modèles et au passage en multimachines du pré-entraînement, ainsi qu'environ 500 heures de calcul GPU correspondant aux 832 *fine-tuning* des modèles lors de l'évaluation, soit un total de 26 000 heures. Le coût environnemental total, selon la documentation du

supercalculateur Jean Zay[4], est équivalent à 6 734 KWh ou 383,84 kg CO2eq, en fonction de l'intensité carbonique du réseau électrique mentionnée par l'étude environnementale BLOOM également réalisée sur Jean Zay (Luccioni et al., 2022).

En ce qui concerne l'évaluation des modèles sur DrBenchmark, approximativement 2 500 heures de calcul GPU V100 ont été utilisés pour réaliser cette étude. Le coût environnemental est estimé à 647,5 KWh ou 36,9 kg CO2eq.

Enfin, pour les travaux sur les modèles Longformer, environ 5 000 heures GPU ont été nécessaires pour pré-entraîner les modèles `DrLongformer-FS` et `DrLongformer-CP` ainsi que 5 000 heures de calcul pour d'autres expériences sur les modèles Longformer dont les résultats se sont avérés décevants et n'ont pas été conservés. Pour l'évaluation de ces modèles sur les différentes tâches, 1 536 *fine-tuning* ont été réalisés correspondant à 1 500 heures GPU. Le coût en *fine-tuning* des modèles Longformer est plus élevé que les modèles BERT. En effet, plus le modèle est large, plus il faudra réduire la taille de lots (*batch size*). Au total, environ 11 500 heures de calcul GPU ont été consommées pour les travaux sur les Longformer, équivalent à 2 978,5 KWh ou 169,77 kg CO2eq.

Le tableau 4.23 résume toutes les consommations relatives à chaque projet ainsi que les consommations des derniers grands modèles de langues à titre de comparaison (Khowaja et al., 2023).

| | **Nombre d'heures GPU** | **Consommation électrique** | **Équivalent kg CO2** |
|---|---|---|---|
| DrBERT | 26 000 | 6 734 KWh | 384 |
| DrBenchmark | 2 500 | 648 KWh | 37 |
| DrLongformer | 11 500 | 2 978 KWh | 170 |
| **Total** | 40 000 | 10 360 KWh | 591 |
| OPT-175B | - | 356 000 KWh | 70 000 |
| BLOOM-175B | - | 475 000 KWh | 25 000 |
| LLaMA-7B | - | 36 000 KWh | 14 000 |
| LLaMA-13B | - | 59 000 KWh | 23 000 |
| LLaMA-33B | - | 233 000 KWh | 90 000 |
| LLaMA-65B | - | 449 000 KWh | 173 000 |
| Total LLaMA | - | 2 638 000 KWh | 300 000 |
| GPT-3-175B | - | 1 287 000 KWh | 500 000 |

TABLEAU 4.23 : Consommation énergétique des travaux d'adaptation de modèle de langue pré-entraînés présentés dans ce manuscrit ainsi que les consommations approximatives de grands modèles de langues récents à titre de comparaison.

Pour une comparaison avec des éléments plus parlants de la vie quotidienne, 591 kg équivalent CO2 correspondent à la consommation de 2 213 litres d'eau en bouteille, à la

4. http://www.idris.fr/media/jean-zay/jean-zay-conso-heure-calcul.pdf

fabrication de 19 smartphones, à un parcours de 250 424 km en TGV, 2 570 km en avion ou 2 716 km en voiture.

## 4.5 Conclusion

Dans ce chapitre, nous avons proposé les premiers modèles de langues pré-entraînés pour les domaines biomédical et clinique : DrBERT et DrLongformer, ainsi que DrBenchmark, un ensemble de jeux de données biomédicales en français permettant de comparer aisément les performances des modèles de langue pré-entraînés.

L'évaluation des modèles fondés sur l'architecture BERT a montré une amélioration de l'état de l'art pour toutes les tâches médicales face aux modèles existants pour le français général (`CamemBERT`) et aux modèles médicaux anglais (`BioBERT`, `PubMedBERT` et `ClinicalBERT`). De plus, nous avons démontré que le pré-entraînement sur une quantité limitée (4 Go) de données biomédicales récupérées sur le web permet de rivaliser avec les modèles pré-entraînés sur des données cliniques.

L'évaluation des modèles reposant sur l'architecture Longformer a montré une amélioration de l'état de l'art pour les tâches de classification de textes par rapport aux modèles BERT existants (`DrBERT` et `CamemBERT-bio`) et aux modèles Longformer anglais `Clinical-Longformer`. Pour les autres tâches que la classification de textes, les résultats sont plus nuancés. Bien que les modèles BERT traitent les tâches de reconnaissance d'entités nommées phrase à phrase, ils restent plus appropriés que les modèles Longformer, peu importe la taille des documents.

À travers ces deux travaux d'adaptation de modèles, nous avons montré qu'il n'y a pas de différences entre le pré-entraînement de zéro et le pré-entraînement poursuivi pour les plus petits modèles comme BERT. En revanche, pour les modèles plus larges, il est préférable d'adapter les modèles en continuant le pré-entraînement d'un modèle existant plutôt que de pré-entraîner un modèle de zéro, comme nous l'avons montré avec les modèles Longformer. Nous avons également démontré qu'il est possible pour la stratégie de pré-entraînement poursuivi de partir d'un modèle anglais spécialisé pour construire un modèle français spécialisé. Cette option est particulièrement intéressante lorsque aucun modèle français n'est disponible pour l'architecture que l'on souhaite adapter. Ces éléments indiquent que le transfert de domaine est réalisable entre deux langues grâce à un pré-entraînement adaptatif à phases multiples.

Chapitre 5

# Extraction d'information en contexte clinique

Dans le chapitre 4, nous avons introduit des modèles de langue adaptés au domaine médical. Ces modèles ont été évalués sur DrBenchmark qui rassemble divers corpus biomédicaux et de cas cliniques. Les résultats ont montré que les modèles spécialisés augmentent les performances état de l'art par rapport aux modèles existants comme CamemBERT. Cependant, ces modèles n'ont pas été évalués dans un contexte clinique réel avec des données issues du soin.

Dans ce chapitre, nous proposons d'évaluer les modèles BERT et Longformer présentés dans le chapitre 4 sur les corpus cliniques présentés dans le chapitre 3 : le corpus du projet GAVROCHE composé d'une tâche de classification de textes et d'une tâche de reconnaissance d'entités nommées et le corpus sur les déterminants sociaux de santé.

## 5.1 Projet GAVROCHE

Dans cette section, nous présentons les expérimentations sur les deux tâches du projet GAVROCHE introduites dans le chapitre 3 : (i) la validation des cas d'insuffisance cardiaque aiguë avec une tâche de classification de texte et (ii) l'extraction des variables d'intérêt du projet avec une tâche de reconnaissance d'entités nommées.

### 5.1.1 Classification de texte - Validation des cas d'ICA

Pour rappel, l'objectif de la tâche de classification de texte est de valider la sélection des CRH extraits de l'EDS pour l'étude en excluant ceux non associés à une hospitalisation pour ICA. Les expériences menées pour cette tâche sont présentées en deux sections. Nous détaillons dans un premier temps le protocole expérimental pour l'affinage des modèles puis, dans un second temps, nous présentons les résultats obtenus par les différents modèles.

#### 5.1.1.1 Protocole expérimental

Nous avons tout d'abord procédé au découpage en jeux d'apprentissage, de validation et de test, tout en veillant à ce que la distribution des classes dans chaque sous-ensemble soit similaire à la distribution initiale du corpus, c'est-à-dire environ 65 % pour les ICA validées et 35 % pour les ICA non validées. Comme indiqué dans le tableau 5.1, 70 % du corpus constitue le jeu d'apprentissage, 10 % le jeu de validation et 20 % le jeu de test.

| | **Apprentissage** | | **Validation** | | **Test** | |
|---|---|---|---|---|---|---|
| **Nb. CR** | 1 179 | | 132 | | 328 | |
| | **Positif** | **Négatif** | **Positif** | **Négatif** | **Positif** | **Négatif** |
| **Nb. CR (%)** | 764 (64,8 %) | 415 (35,2 %) | 86 (65,1 %) | 46 (34,9 %) | 213 (64,9 %) | 115 (35,1 %) |

Tableau 5.1 : Distribution du corpus en jeux d'apprentissage, de validation et de test.

Sept modèles ont été comparés dans cette expérience : `CamemBERT`, `CamemBERT-bio`, `ChuBERT-4GB`, `ChuBERT-Mixed`, `DrBERT-7GB`, `DrBERT-4096` et `DrLongformer-CP`. Les modèles ont été affinés sur la tâche suivant la même configuration que présentée dans les travaux sur les Longformer en Section 4.3. À savoir, les documents sont tronqués à 512 tokens pour les modèles BERT et à 4 096 tokens pour les modèles Longformer. Plusieurs taux d'apprentissage (*learning rate*) ont été expérimentés ($\{1e-4, 1e-5, 2e-5, 5e-5\}$) avec quatre exécutions pour chaque taux d'apprentissage. Nous conservons les résultats

du taux d'apprentissage obtenant la meilleure moyenne sur les quatre exécutions. Les modèles sont évalués sur le jeu de test avec les métriques de précision, rappel et F1 score.

#### 5.1.1.2 Résultats

Les performances des modèles sont présentées dans le tableau 5.2. Pour les modèles BERT, nous observons peu de différences en termes de F1 Score entre le modèle général `CamemBERT` et les modèles biomédicaux avec des gains de 1,79 pour `CamemBERT-bio` et 1,57 pour `DrBERT-7GB`. En revanche, les différences avec les modèles cliniques sont un peu plus élevées avec des gains de 2,48 pour `ChuBERT-4GB` et 3,19 pour `ChuBERT-Mixed`. En comparant les modèles BERT et Longformer, nous observons une différence significative, notamment avec le modèle `DrLongformer-CP`. En effet, ce modèle montre des gains de F1 score allant de 3,95 par rapport à `CamemBERT-bio` à 5,74 par rapport à `CamemBERT`.

| | **Précision** | **Rappel** | **F1 Score** |
|---|---|---|---|
| `CamemBERT` | 81,99 | 81,56 | 81,68 |
| `CamemBERT-bio` | 83,45 | 83,54 | 83,47 |
| `ChuBERT-4GB` | 84,27 | 84,30 | 84,16 |
| `ChuBERT-Mixed` | 85,03 | 84,91 | 84,87 |
| `DrBERT-7GB` | 83,39 | 83,23 | 83,25 |
| `DrBERT-4096` | <u>85,56</u> | <u>85,57</u> | <u>85,22</u> |
| `DrLongformer-CP` | **87,50** | **87,58** | **87,42** |

Tableau 5.2 : Performances moyennes des modèles sur quatre exécutions pour la validation des cas d'insuffisance cardiaque aiguë. Le meilleur modèle est en gras et le second est souligné

Pour chaque modèle, une matrice de confusion de la meilleure exécution parmi les quatre est proposée dans les tableaux 5.3 à 5.9. Les performances des modèles se traduisent par une mauvaise classification de plusieurs dizaines de documents allant de 59 documents mal classés pour `CamemBERT` à 39 documents pour `DrLongformer-CP`. Parmi l'ensemble des documents mal classés par les modèles, 10 documents sont systématiquement mal classés par tous les modèles.

Les Longformers sont les modèles donnant le moins de faux négatifs. Pour la validation des cas d'ICA, il est important d'avoir un nombre de faux négatifs le plus faible possible puisque l'on ne veut pas exclure à tort des patients d'intérêts pour l'étude. Concernant les faux positifs, nous cherchons à les minimiser mais ils sont moins importants que les faux négatifs. L'objectif est d'affiner le codage médical en excluant les patients hospitalisés

| | Classe réelle | |
|---|---|---|
| Classe prédite | Positif | Négatif |
| Positif | 181 | 27 |
| Négatif | 32 | 88 |

TABLEAU 5.3 : `CamemBERT`.

| | Classe réelle | |
|---|---|---|
| Classe prédite | Positif | Négatif |
| Positif | 191 | 27 |
| Négatif | 22 | 88 |

TABLEAU 5.4 : `CamemBERT-bio`.

| | Classe réelle | |
|---|---|---|
| Classe prédite | Positif | Négatif |
| Positif | 189 | 27 |
| Négatif | 24 | 95 |

TABLEAU 5.5 : `ChuBERT-4GB`.

| | Classe réelle | |
|---|---|---|
| Classe prédite | Positif | Négatif |
| Positif | 195 | 28 |
| Négatif | 18 | 87 |

TABLEAU 5.6 : `ChuBERT-Mixed`.

| | Classe réelle | |
|---|---|---|
| Classe prédite | Positif | Négatif |
| Positif | 190 | 30 |
| Négatif | 23 | 85 |

TABLEAU 5.7 : `DrBERT-7GB`.

| | Classe réelle | |
|---|---|---|
| Classe prédite | Positif | Négatif |
| Positif | 198 | 31 |
| Négatif | 15 | 84 |

TABLEAU 5.8 : `DrBERT-4096`.

| | Classe réelle | |
|---|---|---|
| Classe prédite | Positif | Négatif |
| Positif | 198 | 24 |
| Négatif | 15 | 91 |

TABLEAU 5.9 : `DrLongformer-CP`.

pour d'autres causes que l'ICA. Même si cet affinage n'est pas parfait, tant que le système n'exclut pas à tort les patients hospitalisés pour ICA, celui-ci permettra tout de même d'augmenter la qualité des données incluses dans l'étude, peu importe le nombre de faux positifs.

En analysant la distribution de l'attention sur les comptes rendus cliniques pour la classification d'ICA dans la figure 5.1, nous observons que l'attention est concentrée au début du document, au même endroit que les annotations humaines. En effet, les comptes rendus cliniques commencent systématiquement par le motif de l'hospitalisation qui est un marqueur important pour la classification de l'ICA. Cette concentration de l'attention au début des documents explique également pourquoi BERT obtient de bonnes performances sur cette tâche malgré la limitation aux 512 premiers tokens des comptes rendus. Les pics d'attention après 1 500 mots correspondent aux conclusions trouvées à la fin des comptes

rendus, qui récapitulent les éléments importants du cas clinique. Comme il y a peu de documents dépassant cette longueur, ces pics d'attention sont plus prononcés.

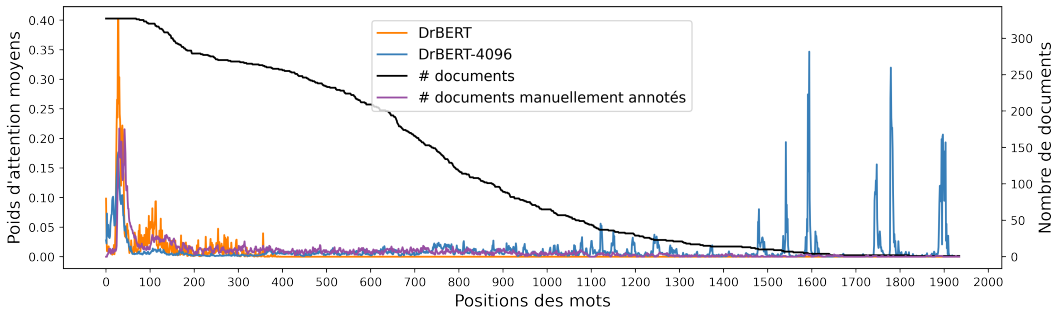


FIGURE 5.1 : Poids moyens d'attention pour chaque position de mot des documents du jeu de test du corpus GAVROCHE. Les poids d'attention sont obtenus à partir du token `[CLS]` utilisé pour la classification en additionnant les poids d'attention de toutes les têtes d'attention de la dernière couche du modèle. Ces poids finaux reflètent l'importance relative de chaque mot, après avoir intégré des informations contextuelles provenant des autres mots dans la phrase à travers les 12 couches d'auto-attention du modèle Transformer. Il est important de noter que ces poids d'attention sont une représentation moyenne, calculée sur l'ensemble des instances de l'ensemble de test, fournissant ainsi une vue d'ensemble de l'importance attribuée à chaque position de mot en moyenne. La tâche de classification d'ICA est annotée tant au niveau du document (Oui/Non) qu'au niveau des mots. Au niveau des mots, des séquences de mots justifiant la classification au niveau du document sont annotées. Cela permet une comparaison entre les annotations humaines et les mécanismes d'attention des modèles.

### 5.1.2 Reconnaissance d'entités nommées - Extraction des variables d'intérêt

Pour rappel, l'objectif de la tâche de reconnaissance d'entités nommées est d'extraire les 24 variables d'intérêt définies par les cliniciens porteurs du projet. Les expériences menées pour cette tâche sont présentées en deux temps. Nous détaillons dans un premier temps le protocole expérimental puis, dans un second temps, nous présentons les résultats obtenus par les différents modèles.

#### 5.1.2.1 Protocole expérimental

Comme indiqué dans le tableau 5.10, les documents du corpus ont été divisés en jeu d'apprentissage (70 %), de validation (10 %) et de test (20 %) en veillant à ce que la

distribution des entités soit équilibrée entre les sous-ensembles (voir tableau 5.11).

| | **Apprentissage** | **Validation** | **Test** |
|---|---|---|---|
| **Nb. CR** | 358 | 51 | 103 |

Tableau 5.10 : Distribution du corpus en jeux d'apprentissage, de validation et de test.

Certaines étiquettes comme l'arrêt cardiaque à l'admission, l'absence de tabagisme et le diabète de type 1 sont présentes dans très peu de CR, ce qui rend difficile l'équilibrage des sous-ensembles.

Les documents de chaque sous-ensemble ont ensuite été formatés au format IOB2 (*inside-outside-beginning*). À l'instar de la tâche de classification, sept modèles ont été comparés dans le cadre de cette expérience : `CamemBERT`, `CamemBERT-bio`, `ChuBERT-4GB`, `ChuBERT-Mixed`, `DrBERT-7GB`, `DrBERT-4096` et `DrLongformer-CP`. Les modèles ont été affinés sur la tâche suivant la même configuration que présentée dans les travaux sur les Longformer en section 4.3. À savoir, les documents sont traités au niveau phrase par les modèles BERT et au niveau document (ou semi-document lorsque le document dépasse les 4 096 tokens) par les modèles Longformer. Plusieurs taux d'apprentissage ont été expérimentés ($\{1e-4, 1e-5, 2e-5, 5e-5\}$) avec quatre exécutions pour chaque taux d'apprentissage. Nous conservons les résultats du taux d'apprentissage obtenant la meilleure moyenne sur les quatre exécutions.

Les modèles sont évalués à deux niveaux sur le jeu de test. La première évaluation est au niveau étiquette avec les métriques de précision, rappel et F1 score de SeqEval en mode strict avec le format IOB2 (*inside-outside-beginning).* La deuxième évaluation est au niveau document en appliquant un ensemble de règles sur les entités prédites afin de produire le jeu de données tabulaires pour l'étude. Cette deuxième évaluation ne s'applique pas aux variables dont l'information est extraite dans le texte, telles que constantes à l'admission (poids, taille, IMC, pression artérielle, fréquence cardiaque) ou les traitements habituels. Pour les autres variables, l'ensemble des règles appliqués aux entités de chaque document permet de gérer les cas où plusieurs étiquettes mutuellement exclusives sont prédites pour une même variable. Ces règles, présentées dans le tableau 5.12, sont appliquées à la fois aux prédictions des modèles et au jeu de données de références.

#### 5.1.2.2 Résultats

Les performances globales des modèles sont présentées dans le tableau 5.13. En premier lieu, il n'y a pas de différences de performances significatives entre les modèles. Le modèle

| **Variable** | **Étiquette** | **Apprentissage** | **Validation** | **Test** |
|---|---|---|---|---|
| Facteur déclenchant ICA | Trouble du rythme | 95 | 10 | 25 |
| | Ischémique | 37 | 0 | 7 |
| | Poussée HTA | 46 | 12 | 14 |
| | Infectieuse | 191 | 37 | 63 |
| | Régime - Traitement | 52 | 2 | 15 |
| | Autre ou inconnu | 172 | 23 | 52 |
| Premier épisode ICA | Oui | 20 | 6 | 5 |
| | Non | 114 | 19 | 21 |
| Traitement habituel | - | 278 | 34 | 67 |
| Arrêt cardiaque admission | Oui | 3 | 0 | 1 |
| Antécédent HTA | Oui | 205 | 33 | 64 |
| Cardiopathie causale | Ischémique | 232 | 20 | 67 |
| | Valvulaire | 225 | 22 | 43 |
| | Rythmique | 162 | 14 | 42 |
| | Autre | 58 | 15 | 20 |
| | Non connue | 11 | 3 | 2 |
| Type d'ICA | Droite isolée | 146 | 14 | 32 |
| | IC gauche décompensée | 486 | 68 | 148 |
| | OAP | 196 | 28 | 62 |
| | Choc cardiogénique | 46 | 10 | 8 |
| Antécédent insuffisance respiratoire chronique | Oui | 52 | 6 | 5 |
| Antécédent BPCO | Oui | 54 | 10 | 15 |
| Antécédent SAOS | Oui | 26 | 4 | 6 |
| Antécédent d'AVC ou AIT | Oui | 62 | 5 | 12 |
| Diabète | Pas de diabète | 5 | 0 | 1 |
| | Diabète de type 1 | 5 | 0 | 3 |
| | Diabète de type 2 | 76 | 11 | 21 |
| | Diabète autre ou indéterminé | 79 | 0 | 18 |
| Fréquence cardiaque à l'admission | - | 241 | 30 | 61 |
| Pression artérielle à l'admission | - | 198 | 26 | 56 |
| Poids | - | 142 | 16 | 34 |
| Taille | - | 57 | 5 | 15 |
| IMC | - | 31 | 4 | 7 |
| Tabagisme | Jamais | 10 | 0 | 2 |
| | Sevré | 45 | 7 | 17 |
| | Actif | 19 | 3 | 7 |
| Antécédent troubles du rythme | Oui | 61 | 13 | 14 |
| Antécédent dépression | Oui | 26 | 6 | 8 |
| Antécédent troubles cognitifs | Oui | 101 | 3 | 36 |
| Antécédent de cancer | Oui | 90 | 18 | 26 |
| AC/FA à l'admission | Oui | 94 | 6 | 29 |
| FEVG | $\leq 40$ | 96 | 13 | 25 |
| | $41 \leq \text{FEVG} \leq 49$ | 18 | 0 | 5 |
| | $\geq 50$ | 98 | 16 | 31 |

Tableau 5.11 : Distribution des étiquettes dans les jeux d'apprentissage, de validation et de test.

| **Variable** (Étiquettes) | **Règle** | **Exemples** |
|---|---|---|
| **Premier épisode ICA**<br>- Oui<br>- Non | On garde l'étiquette Non | Prédiction au niveau document : Oui, Non<br>Étiquette finale après application de la règle : Non |
| **Diabète**<br>- Pas de diabète<br>- Type 1<br>- Type 2<br>- Autre | Si Type 1 et Autre, on garde Autre<br>Si Type 2 et Autre, on garde Autre<br>Si Type 1 et Type 2, on garde Autre | Prédiction au niveau document : Type 1, Autre<br>Étiquette finale après application de la règle : Autre<br>Prédiction au niveau document : Type 1, Type 2<br>Étiquette finale après application de la règle : Autre |
| **FEVG**<br>- $\leq 40$<br>- $41 \leq$ FEVG $\leq 49$<br>- $> 50$ | On garde la FEVG la plus sévère | Prédiction au niveau document : FEVG $\leq 40$ et FEVG $> 50$<br>Étiquette finale après application de la règle : FEVG $\leq 40$ |
| **Tabagisme**<br>- Actif<br>- Sevré<br>- Jamais | On garde l'étiquette suivant l'ordre de priorité :<br>Sevré > Actif > Jamais | Prédiction au niveau document : Sevré et Actif<br>Étiquette finale après application de la règle : Sevré |

Tableau 5.12 : Règles de décision appliquées aux variables ayant des modalités mutuellement exclusives pour générer l'étiquette finale de la variable pour le document.

`DrLongformer-CP` est le meilleur modèle avec un F1 score de 44,58, suivi de très près par le modèle `CamemBERT` avec un F1 score de 44,51.

| **Modèle** | **Précision** | **Rappel** | **F1 Score** |
|---|---|---|---|
| `CamemBERT` | **42,91** | <u>46,34</u> | <u>44,51</u> |
| `CamemBERT-bio` | <u>41,45</u> | 46,54 | 43,84 |
| `ChuBERT-4GB` | 36,20 | 41,52 | 38,66 |
| `ChuBERT-Mixed` | 34,63 | 39,40 | 36,84 |
| `DrBERT-7GB` | 36,66 | 41,19 | 38,75 |
| `DrBERT-4096` | 39,46 | 45,03 | 42,05 |
| `DrLongformer-CP` | 41,35 | **48,38** | **44,58** |

Tableau 5.13 : Performances moyennes des modèles sur quatre exécutions pour l'extraction des variables d'intérêt. Le meilleur modèle est en gras et le second est souligné.

En examinant plus en détails les performances de `DrLongformer-CP` pour chaque entité présentée dans le tableau 5.14, nous observons que celles-ci sont très variables d'une entité à une autre. Les performances sont souvent corrélées avec le nombre d'annotations. En effet, plus il y a d'annotations pour une entité donnée, meilleures sont les performances obtenues. Cette observation est valable pour des entités telles que les traitements habituels, les cardiopathies causales et les types d'ICA gauche décompensée et OAP. À l'inverse, les entités avec peu d'annotations, comme la plupart des facteurs déclenchant d'ICA, la notion d'arrêt cardiaque à l'admission, ainsi que l'absence de tabagisme et de diabète, posent des difficultés au modèle et se traduisent par des scores très faibles.

Cependant, la tendance inverse est aussi observée pour certaines entités, comme les

| **Variable** | **Étiquette** | **Précision** | **Rappel** | **F1 Score** |
|---|---|---|---|---|
| Facteur déclenchant ICA | Trouble du rythme | 3,15 | 2,73 | 2,92 |
| | Ischémique | 0,00 | 0,00 | 0,00 |
| | Poussée HTA | 5,13 | 8,33 | 6,35 |
| | Infectieuse | 33,41 | 33,84 | 33,58 |
| | Régime - Traitement | 5,55 | 1,06 | 1,79 |
| | Autre ou inconnu | 10,03 | 10,53 | 10,18 |
| Premier épisode ICA | Oui | 5,00 | 6,25 | 5,56 |
| | Non | 14,48 | 16,25 | 15,26 |
| Traitement habituel | - | 59,69 | 78,90 | 67,87 |
| Arrêt cardiaque admission | Oui | 0,00 | 0,00 | 0,00 |
| Antécédent HTA | Oui | 70,17 | 83,19 | 76,08 |
| Cardiopathie causale | Ischémique | 39,26 | 60,02 | 47,39 |
| | Valvulaire | 43,00 | 63,11 | 51,13 |
| | Rythmique | 44,13 | 59,15 | 50,53 |
| | Autre | 15,38 | 09,93 | 11,98 |
| | Non connue | 0,00 | 0,00 | 0,00 |
| Type d'ICA | Droite isolée | 17,52 | 20,87 | 19,04 |
| | IC gauche décompensée | 52,71 | 61,32 | 56,68 |
| | OAP | 59,97 | 62,72 | 61,25 |
| | Choc cardiogénique | 20,05 | 12,50 | 15,25 |
| Antécédent insuffisance respiratoire chronique | Oui | 45,21 | 34,38 | 37,35 |
| Antécédent BPCO | Oui | 56,32 | 64,55 | 60,00 |
| Antécédent SAOS | Oui | 45,83 | 33,33 | 37,50 |
| Antécédent d'AVC ou AIT | Oui | 45,53 | 55,47 | 49,83 |
| Diabète | Pas de diabète | 0,00 | 0,00 | 0,00 |
| | Diabète de type 1 | 0,00 | 0,00 | 0,00 |
| | Diabète de type 2 | 61,31 | 81,49 | 69,94 |
| | Diabète autre ou indéterminé | 24,27 | 43,18 | 31,02 |
| Fréquence cardiaque à l'admission | - | 24,73 | 32,50 | 28,08 |
| Pression artérielle à l'admission | - | 33,72 | 41,20 | 37,07 |
| Poids | - | 16,06 | 20.00 | 17,79 |
| Taille | - | 17,09 | 14,06 | 15,35 |
| IMC | - | 7,81 | 8,33 | 7,93 |
| Tabagisme | Jamais | 0,00 | 0,00 | 0,00 |
| | Sevré | 56,20 | 86,16 | 67,86 |
| | Actif | 22,04 | 21,59 | 21,38 |
| Antécédent troubles du rythme | Oui | 0,00 | 0,00 | 0,00 |
| Antécédent dépression | Oui | 32,97 | 18,33 | 23,44 |
| Antécédent troubles cognitifs | Oui | 38,60 | 48,91 | 42,97 |
| Antécédent de cancer | Oui | 46,43 | 56,96 | 51,11 |
| AC/FA à l'admission | Oui | 23,81 | 33,59 | 27,85 |
| FEVG | $\leq 40$ | 28,01 | 35,00 | 31,07 |
| | $41 \leq \text{FEVG} \leq 49$ | 0,00 | 0,00 | 0,00 |
| | $\geq 50$ | 54,21 | 54,31 | 54,16 |

TABLEAU 5.14 : Résultats obtenus pour chaque entité par le meilleur modèle `DrLongformer-CP` sur le jeu de test.

| Variable | Étiquette | Nb. Doc | Spécificité | Sensibilité | Précision | F1 Score |
|---|---|---|---|---|---|---|
| Facteur déclenchant ICA | Trouble du rythme | 14 | 88,76 | 71,43 | 50.00 | 58,82 |
| | Ischémique | 6 | 97,93 | 33,33 | 50,00 | 40,00 |
| | Poussée HTA | 10 | 97,85 | 97,84 | 80,00 | 80,00 |
| | Infectieuse | 30 | 87,67 | 86,67 | 74,29 | 80,00 |
| | Régime - Traitement | 11 | 100,00 | 54,55 | 100,00 | 70,59 |
| | Autre ou inconnu | 25 | 82,05 | 68,00 | 54,84 | 60,71 |
| Premier épisode ICA | Oui | 4 | 98,99 | 0,00 | 0,00 | 0,00 |
| | Non | 16 | 87,36 | 68,75 | 50,00 | 57,89 |
| Arrêt cardiaque admission | Oui | 1 | 100,00 | 0,00 | 0,00 | 0,00 |
| Antécédent HTA | Oui | 54 | 91,84 | 96,30 | 92,86 | 94,55 |
| Cardiopathie causale | Ischémique | 32 | 94,36 | 100,00 | 88,89 | 94,12 |
| | Valvulaire | 26 | 92,21 | 92,31 | 80,00 | 85,71 |
| | Rythmique | 33 | 77,14 | 90,91 | 65,22 | 75,95 |
| | Autre | 12 | 91,21 | 41,67 | 38,46 | 40,00 |
| | Non connue | 2 | 100,00 | 0,00 | 0,00 | 0,00 |
| Type d'ICA | Droite isolée | 17 | 79,07 | 58,82 | 35,71 | 44,44 |
| | IC gauche décompensée | 65 | 63,16 | 92,31 | 81,08 | 86,33 |
| | OAP | 35 | 85,29 | 80,00 | 73,68 | 76,71 |
| | Choc cardiogénique | 5 | 96,93 | 60,00 | 50,00 | 54,55 |
| Antécédent insuffisance respiratoire chronique | Oui | 5 | 97,96 | 40,00 | 50,00 | 44,44 |
| Antécédent BPCO | Oui | 12 | 97,80 | 75,00 | 81,82 | 78,26 |
| Antécédent SAOS | Oui | 5 | 100,00 | 60,00 | 100,00 | 75,00 |
| Antécédent d'AVC ou AIT | Oui | 10 | 97,85 | 100,00 | 83,33 | 90,91 |
| Diabète | Pas de diabète | 1 | 100,00 | 0,00 | 0,00 | 0,00 |
| | Diabète de type 1 | 2 | 100,00 | 0,00 | 0,00 | 0,00 |
| | Diabète de type 2 | 18 | 90,59 | 94,44 | 68,00 | 79,07 |
| | Diabète autre ou indéterminé | 12 | 94,51 | 83,33 | 66,67 | 74,07 |
| Tabagisme | Jamais | 2 | 100,00 | 50,00 | 100,00 | 66,67 |
| | Sevré | 16 | 95,40 | 100,00 | 80,00 | 88,89 |
| | Actif | 7 | 97,92 | 42,86 | 60,00 | 50,00 |
| Antécédent troubles du rythme | Oui | 12 | 98,90 | 41,67 | 83,33 | 55,56 |
| Antécédent dépression | Oui | 7 | 98,96 | 85,71 | 85,71 | 85,71 |
| Antécédent troubles cognitifs | Oui | 15 | 95,45 | 86,67 | 76,47 | 81,25 |
| Antécédent de cancer | Oui | 17 | 93,02 | 88,23 | 71,43 | 78,95 |
| AC/FA à l'admission | Oui | 23 | 85,00 | 86,96 | 62,50 | 72,73 |
| FEVG | $\leq 40$ | 13 | 86,87 | 84,62 | 47,83 | 61,11 |
| | $41 \leq \text{FEVG} \leq 49$ | 3 | 100,00 | 0,00 | 0,00 | 0,00 |
| | $\geq 50$ | 23 | 96,25 | 73,91 | 85,00 | 79,07 |

TABLEAU 5.15 : Résultats obtenus au niveau document pour chaque entité par le meilleur modèle `DrLongformer-CP` sur le jeu de test après application des règles de sélection des modalités.

constantes à l'admission (la fréquence cardiaque, la pression artérielle et le poids) et le type d'ICA droite isolée dont le nombre d'annotations est élevé mais les F1 scores ne dépassent pas 30 %. Nous pouvons faire la même remarque pour les entités « Autre ou inconnu » des facteurs déclenchant de l'ICA et « Non » du premier épisode d'ICA. Ces observations montrent que lorsqu'il y a trop de variété syntaxique dans les annotations d'une même entité, le modèle rencontrera des difficultés.

Enfin, d'autres entités telles que l'antécédent de BPCO, le tabagisme sevré ou le diabète de type 2 affichent des performances entre 60 % et 70 % de F1 score malgré une

faible quantité d'annotations. Les annotations pour ces entités sont souvent exprimées de la même façon dans les comptes rendus, ce qui facilite l'apprentissage du modèle.

Enfin, en appliquant l'ensemble de règles pour générer le jeu de données tabulaires final, nous obtenons les résultats présentés dans le tableau 5.15. Globalement, il y a une différence notable entre les performances des modèles au niveau entités et au niveau document. Pour la plupart des variables, les modèles ne retrouvent pas toutes les entités annotées manuellement mais en identifient seulement un sous-ensemble, ce qui permet tout de même d'avoir de bons résultats au niveau document.

Pour d'autre variables, comme « Facteur déclenchant ICA : Ischémique » ou « Antécédent troubles du rythme », les performances en F1 score au niveau entité sont de 0,00 % contre 40 % et 55,56 % au niveau document. Cette différence de performance indique que les modèles trouvent d'autres entités qui n'ont pas été annotées par les cliniciens.

## 5.2 Déterminants sociaux de santé

Dans cette section, nous présentons les expérimentations sur le corpus de déterminants sociaux introduit dans le chapitre 3. Bien que le corpus soit annoté avec un schéma composé d'entités et de relations, seules les entités ont été traitées dans nos expérimentations. Nous présentons celles-ci en deux temps : le protocole expérimental pour l'affinage des modèles, puis les résultats obtenus par les modèles.

### 5.2.1 Protocole expérimental

À l'instar des autres corpus de reconnaissance d'entités nommées, nous avons converti le corpus de déterminants sociaux de santé au format IOB2. Cependant, ce format ne permet pas de traiter les annotations imbriquées (i.e. les mots ayant plusieurs étiquettes). Les annotations ont été filtrées pour conserver uniquement les annotations de plus large granularité, comme décrit par Touchent et al. (2023). Les documents constituant ce corpus sont courts (150 tokens maximum), aucune troncation n'est donc nécessaire pour les modèles BERT. Les documents du corpus ont ensuite été divisés en jeu d'apprentissage (72 %), de validation (8 %) et de test (20 %), comme indiqué dans le tableau 5.16. La distribution pour chaque entité est présentée dans le tableau 5.17.

| | **Apprentissage** | **Validation** | **Test** |
|---|---|---|---|
| **Nb. CR** | 1 190 | 170 | 340 |

Tableau 5.16 : Distribution des documents du corpus de déterminants sociaux de santé en jeux d'apprentissage, de validation et de test.

Sept modèles ont été affinés et évalués : `CamemBERT`, `CamemBERT-bio`, `ChuBERT-4GB`, `ChuBERT-Mixed`, `DrBERT-7GB`, `DrBERT-4096` et `Drlongformer-CP`. Chaque modèle a été exécuté quatre fois avec un taux d'apprentissage de $2e-5$ et un *early stopping* de cinq *epoch*. La technique d'*early stopping* permet d'éviter le surapprentissage des modèles en interrompant leur optimisation lorsque la fonction de perte sur le jeu de validation ne diminue plus durant un nombre prédéterminé d'itérations. Enfin, les modèles sont évalués sur les entités avec les métriques de précision, rappel et F1 score de SeqEval.

| Entités | Apprentissage | Validation | Test |
|---|---|---|---|
| Activité physique : Oui | 152 | 23 | 33 |
| Activité physique : Non | 38 | 6 | 11 |
| Consommation d'alcool | 436 | 62 | 111 |
| Consommation de drogues | 76 | 11 | 17 |
| Conditions de vie : Cohabitation | 384 | 56 | 121 |
| Conditions de vie : Seul | 193 | 32 | 53 |
| Dernière profession exercée | 249 | 33 | 69 |
| Descendance : Non | 98 | 13 | 25 |
| Descendance : Oui | 774 | 90 | 204 |
| Domicile : Oui | 465 | 54 | 137 |
| Domicile : Non | 16 | 3 | 3 |
| Niveau d'éducation | 56 | 5 | 17 |
| Origine ethnique | 69 | 13 | 21 |
| Profession | 445 | 63 | 147 |
| Revenu | 24 | 3 | 13 |
| Statut matrimonial : Célibataire | 50 | 8 | 14 |
| Statut matrimonial : Divorcé | 69 | 13 | 13 |
| Statut matrimonial : Marié | 328 | 35 | 100 |
| Statut matrimonial : Veuf | 69 | 8 | 17 |
| Statut d'emploi : Actif | 221 | 34 | 53 |
| Statut d'emploi : Autre | 81 | 3 | 32 |
| Statut d'emploi : Étudiant | 18 | 4 | 12 |
| Statut d'emploi : Retraité | 291 | 42 | 86 |
| Statut d'emploi : Sans emploi | 112 | 18 | 31 |
| Tabagisme | 608 | 88 | 159 |
| **Entités pour les relations** | **Apprentissage** | **Validation** | **Test** |
| Antécédent | 253 | 31 | 73 |
| Durée | 48 | 7 | 18 |
| Fréquence | 414 | 52 | 98 |
| Quantité | 654 | 92 | 175 |
| StatusTime : Actuel | 255 | 51 | 82 |
| StatusTime : Aucun | 478 | 69 | 127 |
| StatusTime : Passé | 238 | 27 | 55 |
| Type | 187 | 24 | 46 |

Tableau 5.17 : Distribution des entités dans les jeux d'apprentissage, de validation et de test du corpus de déterminants sociaux de santé.

### 5.2.2 Résultats

Les performances globales des modèles sont présentées dans le tableau 5.18. `DrLongformer-CP` montre les meilleures performances avec un F1 score de 62,61 %. Les autres modèles obtiennent des scores assez proches, avec des F1 score allant de 62,13 % pour `DrBERT-4096` à 57,71 % pour `ChuBERT-4GB`, excepté pour le modèle `CamemBERT` qui rencontre des difficultés lors de l'apprentissage avec un F1 score de 30,85 %. En effet, le *early stopping* met fin à l'affinage du modèle à la 5ème *epoch* car il n'a pas réussi à prédire une seule entité correcte. Cela montre qu'un modèle non spécialisé a besoin d'un apprentissage plus long pour obtenir des performances comparables aux modèles spécialisés.

Bien que ce corpus soit composé de documents cliniques, l'usage d'un modèle clinique comme `ChuBERT-4GB` ou `ChuBERT-Mixed` n'apporte pas de gain de performances par rapport aux modèles biomédicaux. Les déterminants sociaux de santé traitent des concepts généraux de la vie quotidienne tels que la profession, l'éducation, la consommation de tabac et d'alcool. Ces concepts peuvent donc être facilement modélisés par les modèles biomédicaux

| **Modèle** | **Précision** | **Rappel** | **F1 Score** |
|---|---|---|---|
| `CamemBERT` | 29,98 | 31,79 | 30,85 |
| `CamemBERT-bio` | 60,17 | 63,28 | 61,68 |
| `ChuBERT-4GB` | 54,89 | 58,68 | 57,71 |
| `ChuBERT-Mixed` | 55,97 | 60,47 | 58,12 |
| `DrBERT-7GB` | 60,23 | 63,74 | 61,93 |
| `DrBERT-4096` | <u>60,42</u> | <u>64,00</u> | <u>62,13</u> |
| `DrLongformer-CP` | **61,21** | **64,09** | **62,61** |

Tableau 5.18 : Performances moyennes des modèles sur quatre exécutions pour la reconnaissance d'entités nommées du corpus sur les déterminants sociaux. Le meilleur modèle est en gras et le second est souligné.

Pour les performances par entité du meilleur modèle `DrLongformer-CP` présenté dans le tableau 5.19, nous observons les mêmes tendances que pour la tâche de reconnaissance d'entités nommées du projet GAVROCHE avec une corrélation entre les performances du modèle et la variabilité syntaxique des annotations par entité. En effet, les entités pour lesquelles l'information est toujours exprimée de la même façon, comme « Descendance : Non », « Statut matrimonial : Célibataire », sont facilement apprises par le modèle. À l'inverse, les entités comme « Descendance : Oui », « Domicile : Oui », « Profession » et « Dernière profession exercée » posent plus de difficulté au modèle, malgré une grande

| **Entités** | **Précision** | **Rappel** | **F1 Score** |
|---|---|---|---|
| Activité physique : Non | 17,65 | 15,00 | 16,22 |
| Activité physique : Oui | 30,12 | 32,89 | 31,45 |
| Consommation d'alcool | 68,23 | 73,12 | 70,59 |
| Consommation de drogues | 82,05 | 88,89 | 85,33 |
| Conditions de vie : Cohabitation | 76,26 | 73,61 | 74,91 |
| Conditions de vie : Seul | 75,00 | 89,06 | 81,43 |
| Dernière profession exercée | 28,30 | 17,86 | 21,90 |
| Descendance : Non | 82,76 | 96,00 | 88,89 |
| Descendance : Oui | 44,20 | 42,05 | 43,10 |
| Domicile : Oui | 36,64 | 45,35 | 40,53 |
| Domicile : Non | 0,00 | 0,00 | 0,00 |
| Niveau d'éducation | 29,55 | 44,83 | 35,61 |
| Origine ethnique | 83,02 | 75,86 | 79,28 |
| Profession | 48,12 | 58,38 | 52,75 |
| Revenu | 78,95 | 78,95 | 78,95 |
| Statut matrimonial : Célibataire | 100,00 | 100,00 | 100,00 |
| Statut matrimonial : Divorcé | 81,08 | 85,71 | 83,33 |
| Statut matrimonial : Marié | 73,33 | 77,10 | 75,17 |
| Statut matrimonial : Veuf | 96,97 | 94,12 | 95,52 |
| Statut d'emploi : Actif | 54,07 | 76,00 | 63,18 |
| Statut d'emploi : Autre | 36,84 | 38,18 | 37,50 |
| Statut d'emploi : Étudiant | 78,95 | 65,22 | 71,43 |
| Statut d'emploi : Retraité | 81,13 | 69,35 | 74,78 |
| Statut d'emploi : Sans emploi | 45,65 | 42,00 | 43,75 |
| Tabagisme | 94,12 | 97,60 | 95,83 |
| **Entités pour les relations** | **Précision** | **Rappel** | **F1 Score** |
| Antécédent | 61,57 | 79,29 | 69,32 |
| Durée | 70,37 | 79,17 | 74,51 |
| Fréquence | 58,14 | 70,42 | 63,69 |
| Quantité | 46,15 | 48,34 | 47,22 |
| StatusTime : Actuel | 58,26 | 40,12 | 47,52 |
| StatusTime : Aucun | 66,11 | 70,83 | 68,39 |
| StatusTime : Passé | 93,04 | 91,30 | 92,16 |
| Type | 54,44 | 44,55 | 49,00 |

Tableau 5.19 : Distribution des entités dans les jeux d'apprentissage, de validation et de test du corpus de déterminants sociaux de santé.

quantité d'annotations.

## 5.3 Conclusion

Dans ce chapitre, nous avons proposé une évaluation en contexte clinique des modèles de langues pré-entraînés pour les domaines biomédical et clinique introduits dans le chapitre 4. L'évaluation des modèles a montré une amélioration de l'état de l'art et une plus grande robustesse par rapport au modèle `CamemBERT` pour le français général. De plus, nous avons montré que les modèles de langue pré-entraînés sur des données biomédicales récoltées sur le web peuvent rivaliser avec les modèles de langue pré-entrainés sur des données cliniques.

Enfin, certains résultats peuvent être considérés comme satisfaisants dans le domaine du TAL mais ce n'est pas forcément le cas dans le domaine médical où les attentes sont plus élevées envers les performances des modèles pour une utilisation en recherche clinique.

# Conclusion

Dans cette thèse, nous nous sommes intéressés à l'application du Traitement Automatique du Langage (TAL) dans le domaine médical, et ce, à travers diverses thématiques : la création de corpus, l'adaptation de modèles au domaine et l'évaluation extrinsèque des systèmes. En travaillant dans ce domaine, l'une des problématiques majeures que nous avons identifiées est le manque de ressources disponibles en français pour le domaine médical. En effet, malgré un domaine en plein essor ces dernières années, peu de corpus sont disponibles en libre accès, et parmi ceux-ci, aucun n'inclut des données cliniques issues des soins aux patients.

Parmi les autres ressources manquantes, les modèles de langue pré-entraînés sont également à compter. Seuls des modèles généralistes tels que `CamemBERT` et `FlauBERT` sont disponibles et utilisés par la communauté, indépendamment du domaine de spécialité. Pourtant, des travaux dans d'autres langues ont démontré l'intérêt d'adapter les modèles au domaine spécifique pour améliorer les performances (El Boukkouri et al., 2022; Lee et al., 2019; Chalkidis et al., 2020; Li et al., 2023).

La technicité de la langue médicale requiert des modèles adaptés à ce domaine pour exploiter au mieux les données cliniques. Nous nous sommes donc concentrés sur l'adaptation de modèles de langue pré-entraînés au domaine médical et la création de nouveaux corpus pour enrichir le champ des corpus existants.

Pour conclure, nous résumons les principales contributions présentées dans cette thèse avant de proposer des perspectives pour des recherches futures.

## 5.4 Contributions

Nos principales contributions portent sur une revue systématique de la littérature sur l'usage du TAL dans les entrepôts de données de santé, l'adaptation des modèles BERT et Longformer au domaine médical français, ainsi qu'une évaluation de ces modèles sur les tâches de DrBenchmark — un benchmark de corpus biomédicaux constitué pour cette occasion — et sur des corpus cliniques composés de comptes rendus hospitaliers du CHU de Nantes, annotés durant cette thèse.

Notre première contribution est une revue systématique de la littérature sur l'usage du TAL dans les entrepôts de données de santé. Notre objectif était d'identifier les tâches de TAL qui sont employées pour traiter les documents textuels issus des EDS ainsi que les méthodes utilisées pour traiter ces tâches. Nos résultats montrent que les méthodes de TAL sont principalement utilisées pour extraire des informations renseignées dans les documents textuels stockés dans les EDS et pour identifier des populations de patients. En fonction de la tâche cible, différentes méthodes sont utilisées, allant des méthodes symboliques aux méthodes d'apprentissage automatique et d'apprentissage profond pour les plus récentes. Les méthodes symboliques et linguistiques sont encore largement utilisées ces dernières années pour réaliser la plupart des tâches, notamment grâce à des approches à base d'expressions régulières ou la recherche de motifs en exploitant des lexiques spécialisés, comme des listes de médicaments ou des terminologies. Nos résultats indiquent également que peu de langues sont représentées dans cette revue systématique, avec notamment une sur-représentation des travaux sur la langue anglaise. Enfin, cette étude a été réalisée en 2022 et cela a évolué depuis, notamment en France où le développement des EDS a été très rapide ces dernières années. Cette revue systématique de la littérature a fait l'objet de deux publications. La première version de cette étude a été publiée dans la conférence nationale TALN-RECITAL (Bazoge, 2021), puis, une deuxième version plus étoffée a été publié dans le journal JMIR Medical Informatics (Bazoge et al., 2023).

Notre deuxième contribution est la constitution de DrBenchmark, un benchmark de tâches biomédicales françaises permettant d'évaluer les modèles de langue pré-entraînés. Ce benchmark rassemble la majorité des corpus biomédicaux français introduits précédemment dans la littérature scientifique. DrBenchmark assure une comparaison équitable entre les systèmes grâce à des scripts de pré-entraînement et d'évaluation dédiés à chacune des tâches. Les corpus inclus dans DrBenchmark sont formatés suivants des schémas normalisés et des sous-ensembles de données prédéfinies grâce à la bibliothèque Datasets de HuggingFace (Lhoest et al., 2021). Pour cette contribution, un article a été soumis pour publication à la conférence LREC-COLING 2024.

Notre troisième contribution est l'adaptation des modèles de langue BERT et Longformer aux domaines biomédical et clinique français. Nous avons tout d'abord introduit `DrBERT` et `ChuBERT`, deux modèles de langues pré-entraînés, le premier sur des données

biomédicales récoltées sur le web et le second sur des phrases simples extraites de comptes rendus hospitaliers du CHU de Nantes. Au travers d'une étude comparative, nous avons étudié l'impact de différents paramètres sur les performances des modèles de langues pré-entraînés : les stratégies de pré-entraînement de modèles (pré-entraînement de zéro et pré-entraînement poursuivi) et les sources de données (les données biomédicales récoltées sur le web et les données cliniques issues de l'entrepôt de données du CHU de Nantes). Nos résultats ont montré que les modèles spécialisés apportent une amélioration de l'état de l'art pour toutes les tâches médicales face aux modèles existants pour le français général (`CamemBERT`) et aux modèles médicaux anglais (`BioBERT`, `PubMedBERT` et `ClinicalBERT`). Nous avons également montré qu'à quantité de données limitées (4 Go), les modèles pré-entraînés sur des données médicales récupérées sur le web permettent de rivaliser avec les modèles pré-entraînés sur des données cliniques. De plus, nous avons montré qu'il n'y a pas de différences entre le pré-entraînement de zéro et le pré-entraînement poursuivi pour les modèles BERT. Ces travaux ont fait l'objet d'une publication à la conférence *Annual Meeting of the Association for Computational Linguistics* (ACL) pour laquelle nous avons été récompensés d'une mention honorable (Labrak et al., 2023a).

Nous avons ensuite introduit DrLongformer, un modèle pré-entraîné sur des données biomédicales françaises permettant de traiter des séquences de textes allant jusqu'à 4 096 tokens. Dans ce travail d'adaptation des modèles Longformer, notre objectif était d'évaluer à nouveau les stratégies de pré-entraînement des modèles (pré-entraînement de zéro, pré-entraînement poursuivi et conversion d'un modèle BERT en modèle Longformer) et de vérifier si les résultats observés lors de l'adaptation des modèles BERT étaient de nouveau observés avec les modèles Longformer. L'évaluation des modèles a montré une amélioration de l'état de l'art pour les tâches de classification de textes par rapport aux modèles BERT existants (`DrBERT` et `CamemBERT-bio`) et aux modèles Longformer anglais `Clinical-Longformer`. Pour les autres tâches que la classification de textes, les résultats sont plus nuancés. Bien que les modèles BERT traitent les tâches de reconnaissance d'entités nommées phrase à phrase, ils restent plus appropriés que les modèles Longformer, peu importe la taille des documents. Contrairement aux résultats observés avec les modèles BERT, nos résultats avec les modèles Longformer indiquent qu'il est préférable d'adapter les modèles en continuant le pré-entraînement d'un modèle existant plutôt que de pré-entraîner un modèle de zéro, ce qui corrobore les résultats présentés par (El Boukkouri et al., 2022). De plus, nous avons montré qu'il est possible de poursuivre le pré-entraînement d'un modèle anglais spécialisé sur des données biomédicales françaises pour

construire un modèle français spécialisé. Cette option est particulièrement intéressante lorsque aucun modèle français n'est disponible pour l'architecture que l'on souhaite adapter. Ces éléments indiquent que le transfert de domaine est réalisable entre deux langues grâce à un pré-entraînement adaptatif à phases multiples. Enfin, un article a été soumis à la conférence LREC-COLING 2024 pour ces travaux.

Notre quatrième contribution est la création de deux corpus composés de données cliniques issues de l'entrepôt de données de santé du CHU de Nantes. Ces deux corpus s'intègrent dans un contexte de recherche clinique pour deux projets ayant pour but d'extraire des informations sur deux thématiques identifiées dans des comptes-rendus hospitaliers. Le premier corpus est celui du projet GAVROCHE qui a été annoté avec des informations cliniques afin de phénotyper les patients hospitalisés pour insuffisance cardiaque aiguë. Le deuxième corpus porte sur les déterminants sociaux de santé et est composé de paragraphes issus des sections « Mode de vie » des comptes rendus hospitaliers. Ces deux corpus ont permis de compléter l'évaluation des modèles de langue introduits dans cette thèse sur DrBenchmark en apportant une évaluation en domaine clinique qui n'est pas présent dans le benchmark. Nos résultats ont montré que les modèles de langue spécialisés améliorent l'état de l'art et sont plus robustes par rapport au modèle généraliste `CamemBERT`. Enfin, bien que certains résultats puissent être considérés comme suffisants dans le domaine du TAL, ce n'est pas forcément le cas dans le domaine médical où les attentes sont plus élevées envers les performances des modèles pour une utilisation systématique dans la recherche clinique.

## 5.5 Perspectives

Nos contributions ont permis d'apporter de nouvelles ressources pour mieux aborder le domaine médical, que ce soit par la publication de modèles de langues, la création de corpus ou la constitution d'un benchmark d'évaluation.

Les travaux menés durant cette thèse peuvent être poursuivis dans différentes directions, dont quatre nous semble à privilégier. La première concerne la création de corpus. Dans ce manuscrit, nous avons vu qu'il y a peu de corpus médicaux dans la littérature, et ceux-ci partagent souvent des caractéristiques communes, comme les sources de données utilisées (articles scientifiques, cas cliniques) et les tâches traitées (très souvent des tâches de reconnaissance d'entités nommées). Pour pallier ce manque de variété, nous avons introduit plusieurs corpus au cours de cette thèse : les corpus du projet GAVROCHE et

des déterminants sociaux de santé présentés dans le chapitre 3 ainsi que le corpus FrenchMedMCQA composé de questions à choix multiples tirées d'examens de pharmacie et le corpus DiaMed constitué de cas cliniques et annoté avec les chapitres de la CIM-10. La disponibilité de corpus variés et de qualité est nécessaire pour le bon développement des approches de TAL, que ce soit pour l'entraînement des systèmes ou leur évaluation. Beaucoup de tâches ne sont pas encore couvertes dans les corpus du domaine médical français, telles que l'extraction de relations ou l'inférence textuelle (*Natural Language Inference*). De plus, les tâches existantes sont souvent éloignées des besoins cliniques habituels. Il est donc important de continuer à proposer des nouveaux corpus avec des tâches et des sources de données plus variées et plus proches des besoins cliniques.

La deuxième direction de recherche porte sur le partage des données cliniques pour la recherche. De nombreux corpus cliniques et modèles de langue ont été créés dans les établissements hospitaliers et sont conservés en interne, faute de solutions pour pouvoir les partager sans mettre en danger la confidentialité des patients. Nous pensons qu'il serait intéressant de proposer des solutions permettant de partager les corpus et les modèles tout en respectant les contraintes réglementaires et la confidentialité des patients. Parmi les pistes à explorer pour les corpus, nous pensons à l'anonymisation des comptes-rendus en appliquant les techniques détaillées dans l'article G29 sur la protection des données ou à la création de données synthétiques. Ces deux pistes pourraient également servir à pré-entraîner des modèles anonymes partageables. Pour le partage des modèles, la protection de la confidentialité des patients peut être réalisé en aval du pré-entraînement. Par exemple, pour les modèles génératifs, il est possible de bloquer certaines requêtes intrusives ou portant sur des thématiques sensibles.

La troisième direction de recherche concerne l'adaptation des modèles de langue au domaine médicale français. Dans ce manuscrit, nous avons adapté des modèles BERT et Longformer, tous deux composés d'un nombre de paramètres entre 100 et 150 millions. Or, la tendance actuelle est aux grands modèles de langues (*Large Language Models - LLM*) composés de plusieurs milliards de paramètres. La plupart des grands modèles de langue actuels (ChatGPT, BARD, etc.) propose des services en ligne de chat et de questions-réponses, mais cela intègre de nombreux risques à gérer (Bommasani et al., 2022). Des préoccupations ont été soulevées concernant les données utilisées pour l'entraînement de ces modèles privés, ainsi que sur la manière dont les requêtes contenant des données sensibles liées à la vie privée sont transmises sur Internet et la façon dont ces entreprises gèrent ces données (et les utilisent). Ces modèles privés ne peuvent donc pas être utilisés

sur les données cliniques des EDS qui sont soumises au RGPD et au référentiel entrepôt de données de santé. La solution pour adapter et partager un grand modèle de langue médicale français repose de nouveau sur les données ouvertes disponibles sur le web. Cela nécessitera de créer des corpus de qualité adaptés à ces modèles qui sont actuellement manquants pour le français médical, ce qui fait écho aux deux directions de recherche discutées précédemment.

Enfin, la quatrième et dernière direction de recherche concerne l'utilisation des modèles de langue en contexte clinique. En effet, les modèles de langue sont généralement évalués sur des tâches générales issues de benchmarks spécialement conçus à cet effet. Toutefois, ces tâches d'évaluation s'éloignent des besoins cliniques réels, tels que le codage médical ou l'extraction d'informations cliniques complexes, et ne permettent pas de déterminer avec précision si un modèle sera performant dans un cas d'usage très spécifique et spécialisé. Comme nous l'avons observé dans ce manuscrit à travers le projet GAVROCHE, les modèles de langue éprouvent des difficultés avec les tâches complexes nécessitant une expertise clinique approfondie. De plus, ces approches nécessitent un investissement considérable en termes de temps humain d'annotation, pour aboutir finalement à des performances qui ne satisfont pas les attentes des cliniciens. Il serait donc intéressant d'identifier les causes de ces difficultés rencontrées par les modèles. Sont-elles liées à la qualité des données cliniques ? À la façon d'aborder les tâches ? Ou à l'architecture des modèles ? Ce sont autant de pistes qui devraient venir alimenter les travaux de la communauté à l'avenir.

Chapitre 6

# Annexes

## 6.1 Blocs d'annotation du projet GAVROCHE

| **Bloc 1 - Facteur déclenchant, 1er épisode ICA et traitement (11 étiquettes)** | | |
|---|---|---|
| **Variable** | **Définition** | **Étiquettes** |
| Facteur déclenchant de l'ICA | Information souhaitée : L'ICA est-elle liée au facteur déclenchant suivant ? Plusieurs facteurs déclenchant peuvent être indiqués pour un même CRH. | Trouble du rythme |
| | | Ischémique |
| | | Poussée hypertensive |
| | | Infectieux |
| | | Valve |
| | | Régime ou traitement |
| | | Autre ou inconnu |
| Premier épisode ICA | Information souhaitée : est-ce le premier épisode d'ICA connu pour ce patient ? | Oui |
| | | Non |
| Traitement habituel | Cette variable vise à sélectionner l'ensemble des médicaments habituels du patient juste au temps de son admission. | Oui |
| Arrêt cardiaque à l'admission | Information souhaitée : le patient a-t-il été pris en charge en arrêt cardio-respiratoire, et/ou a-t-il présenté un arrêt cardio-respiratoire lors de son transfert au CHU et/ou aux urgences ? | Oui |

Tableau 6.1 : Bloc 1 - Facteur déclenchant de l'ICA, 1er épisode d'ICA, traitement habituel, arrêt cardiaque à l'admission

| **Bloc 2 - Cardiopathie causale et type d'ICA (10 étiquettes)** | | |
|---|---|---|
| **Variable** | **Définition** | **Étiquettes** |
| Antécédent d'hypertension artérielle (HTA) | Information souhaitée : le patient est-il connu comme ayant une hypertension artérielle, contrôle ou non, au temps de l'hospitalisation ? | Oui |
| Cardiopathie causale | Information souhaitée : quel est le facteur étiologique de la cardiopathie causale ? | Ischémique |
| | | Valvulaire |
| | | Rythmique |
| | | Autre |
| | | Non connu |
| Type d'ICA | Correspond aux 4 présentations cliniques de l'ICA d'après les recommendations ESC 2021 | Droite isolée |
| | | Insuffisance cardiaque gauche décompensée |
| | | OAP |
| | | Choc cardiogénique |

Tableau 6.2 : Bloc 2 - Cardiopathie causale et type d'ICA

| **Bloc 3 - Antécédents respiratoires, AVC et diabète (8 étiquettes)** | | |
|---|---|---|
| **Variable** | **Définition** | **Étiquettes** |
| Antécédent d'insuffisance respiratoire chronique | Information souhaitée : le patient était-il connu pour présenter une insuffisance respiratoire chronique (traitée ou non, équilibrée ou non) avant son hospitalisation ? | Oui |
| Antécédent BPCO | Information souhaitée : le patient a-t-il un antécédent de bronchopneumopathie chronique obstructive ? | Oui |
| Antécédent SAOS | Information souhaitée : le patient s'est-il vu diagnostiquer un syndrome d'apnée du sommeil (obstructive ou non) ? | Oui |
| Antécédent d'AVC ou d'AIT | Information souhaitée : le patient a-t-il déjà présenté un accident vasculaire cérébral ? | Oui |
| Diabète | Information souhaitée : le patient est-il connu pour présenter un diabète ? | Pas de diabète |
| | | Diabète de type 1 |
| | | Diabète de type 2 |
| | | Diabète autre ou indéterminé |

Tableau 6.3 : Bloc 3 - Antécédents respiratoires, AVC et diabète

| **Bloc 4 - Infos cliniques à l'admission et tabagisme (8 étiquettes)** | | |
|---|---|---|
| **Variable** | **Définition** | **Étiquettes** |
| Fréquence cardiaque à l'admission | Information souhaitée : quelle est la première fréquence cardiaque (en battements par minute) mesurée lors de l'admission du patient ? | Fréquence cardiaque |
| Pression artérielle à l'admission | Information souhaitée : quelle est la première pression artérielle (systolique et diastolique, en mmHg) mesurée lors l'admission du patient ? | Pression artérielle |
| Poids | Information souhaitée : quel est le poids du patient ? Tous les poids indiqués dans le CRH sont annotés, que ce soit le poids à l'admission, au cours du séjour ou de sortie | Poids |
| Taille | Information souhaitée : quelle est la taille du patient ? | Taille |
| IMC | Information souhaitée : quelle est l'IMC du patient ? | IMC |
| Tabagisme | Information souhaitée : quel est le statut tabagique connu pour le patient lors du temps d'hospitalisation ? | Jamais |
| | | Actif |
| | | Sevré |

Tableau 6.4 : Bloc 4 - Antécédents respiratoires, AVC et diabète

| **Bloc 5 - Antécédent ACFA, 1ère FEVG et autres comorbidités (8 étiquettes)** | | |
|---|---|---|
| **Variable** | **Définition** | **Étiquettes** |
| Antécédent trouble du rythme | Information souhaitée : le patient a-t-il un antécédent d'arythmie cardiaque supraventriculaire/de fibrillation atriale connu avant l'hospitalisation ? | Oui |
| Antécédent de dépression | Information souhaitée : le patient a-t-il un antécédent de dépression ? | Oui |
| Antécédent de troubles cognitifs | Information souhaitée : le patient présente-t-il un trouble cognitif antérieur à l'hospitalisation (même débutant) ? | Oui |
| Antécédent de cancer | Information souhaitée : le patient a-t-il un antécédent de cancer ? On considère ici indifféremment tout cancer solide ou hémopathie, correspondant aux codes ICD-10 débutant par « C » ou de D37 à D48. | Oui |
| AC/FA à l'admission | Information souhaitée : le patient présente-t-il une arythmie cardiaque supraventriculaire/fibrillation atriale au cours de l'admission (au SAMU ou aux urgences, notamment) ? | Oui |
| FEVG | Information souhaitée : quelle est la valeur de la fraction d'éjection ventriculaire gauche évaluée au cours de l'hospitalisation ? | $\leq 40$ |
| | | $41 \leq \text{FEVG} \leq 49$ |
| | | $\geq 50$ |

Tableau 6.5 : Bloc 5 - Antécédents respiratoires, AVC et diabète

## 6.2 CAS corpus

**Étiquetage morpho-syntaxique :** VER:ppre, VER:infi, VER:impf, VER:simp, PUN, DET:POS, ADV, DET:ART, PRO:DEM, INT, VER:futu, VER:subp, VER:cond, VER:pper, KON, NAM, PRO:IND, VER:con, PRP, SYM, SENT, PUN:cit, VER:pres, PRP:det, PRO:REL, PRO:PER, VER:subi, ADJ, NUM, NOM, ABR.

**Classification multi-classes :** negation_speculation, speculation, neutral, negation

**Reconnaissance d'entités nommées - Speculation :** O, B_xcope_inc, I_xcope_inc

**Reconnaissance d'entités nommées - Negation :** O, B_scope_neg, I_scope_neg

## 6.3 ESSAI

**Étiquetage morpho-syntaxique :** INT, PRO:POS, PRP, SENT, PRO, ABR, VER:pres, KON, SYM, DET:POS, VER:, PRO:IND, NAM, ADV, PRO:DEM, NN, PRO:PER, VER:pper, VER:ppre, PUN, VER:simp, PREF, NUM, VER:futu, NOM, VER:impf, VER:subp, VER:infi, DET:ART, PUN:cit, ADJ, PRP:det, PRO:REL, VER:cond, VER:subi.

**Classification multi-classes :** negation_speculation, speculation, neutral, negation.

**Reconnaissance d'entités nommées - Speculation :** O, B_cue_spec, B_scope_spec, I_scope_spec.

**Reconnaissance d'entités nommées - Negation :** O, B_cue_neg, B_scope_neg, I_scope_neg.

## 6.4 QUAERO

GEOG, PHEN, DISO, ANAT, OBJC, PHYS, PROC, DEVI, CHEM, LIVB

## 6.5 E3C

**Clinique :** O, B-CLINENTITY and I-CLINENTITY

**Temporal :** O, B-EVENT, B-ACTOR, B-BODYPART, B-TIMEX3, B-RML, I-EVENT, I-ACTOR, I-BODYPART, I-TIMEX3, I-RML

## 6.6 MorFITT

microbiology, etiology, virology, physiology, immunology, parasitology, genetics, chemistry, veterinary, surgery, pharmacology, psychology

## 6.7 MantraGSC

**Reconnaissance d'entités nommées - Medline :** ANAT, PROC, CHEM, PHYS, GEOG, DEVI, LIVB, OBJC, DISO, PHEN, O.

**Reconnaissance d'entités nommées - EMEA et Patents :** ANAT, PROC, CHEM, PHYS, DEVI, LIVB, OBJC, DISO, PHEN, O.

## 6.8 DEFT-2019 corpus

**Reconnaissance d'entités nommées :** : age, genre, issue and origine.

## 6.9 DEFT-2021

**Classification Multi-étiquettes :** immunitaire (immunology), endocriniennes (endocrinology), blessures (injury), chimiques (chemicals), etatsosy (signs and symptoms), nutritionnelles (nutrition), infections (infections), virales (virology), parasitaires (parasitology), tumeur (oncology), osteomusculaires (osteomuscular disorders), stomatognathique (stomatology), digestif (digestive system disorders), respiratoire (respiratory system disorders), ORL (otorhinolaryngologic diseases), nerveux (nervous system disorders), oeil (eye diseases),

homme (male genital diseases), femme (female genital diseases), cardiovasculaires (cardiology), hemopathies (hemic and lymphatic diseases), genetique (genertic disorders), peau (dermatology).

**Reconnaissance d'entités nommées :** O, B-ANATOMY, I-ANATOMY, B-DATE, I-DATE, B-DOSAGE, I-DOSAGE, B-DURATION, I-DURATION, B-MEDICAL EXAM, I-MEDICAL EXAM, B-FREQUENCY, I-FREQUENCY, B-MODE, I-MODE, B-MOMENT, I-MOMENT, B-PATHOLOGY, I-PATHOLOGY, B-sosy, I-sosy, B-SUBSTANCE, I-SUBSTANCE, B-TREATMENT, I-TREATMENT, B-VALUE, I-VALUE

## 6.10 PxCorpus

**Classification d'intention :** MEDICAL PRESCRIPTION, NEGATE, NONE, REPLACE

**Reconnaissance d'entités nommées :** O, A, CMA_EVENT, D_DOS_FORM, D_DOS_UP, D_DOS_VAL, D_DOS_FORM_EXT, DOS_COND, DOS_UF, DOS_VAL, DRUG, DUR_UT, DUR_VAL, FASTING, FREQ_DAYS, FREQ_INT_V1, FREQ_INT_V1_UT, FREQ_INT_V2, FREQ_INT_V2_UT, FREQ_STARTDAY, FREQ_UT, FREQ_VAL, INN, MAX_UNIT_UF, MAX_UNIT_UT, MAX_UNIT_VAL, MIN_GAP_UT, MIN_GAP_VAL, QSP_UT, QSP_VAL, RE_UT, RE_VAL, RHYTHM_HOUR, RHYTHM_PERDAY, RHYTHM_REC_UT, RHYTHM_REC_VAL, RHYTHM_TDTE, ROA

## 6.11 DiaMed

- A00-B99 *Certain infectious and parasitic diseases*
- C00-D49 *Neoplasms*
- D50-D89 *Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism*
- E00-E89 *Endocrine, nutritional and metabolic diseases*
- F01-F99 *Mental, Behavioral and Neurodevelopmental disorders*
- G00-G99 *Diseases of the nervous system*
- H00-H59 *Diseases of the eye and adnexa*

- H60-H95 *Diseases of the ear and mastoid process*
- I00-I99 *Diseases of the circulatory system*
- J00-J99 *Diseases of the respiratory system*
- K00-K95 *Diseases of the digestive system*
- L00-L99 *Diseases of the skin and subcutaneous tissue*
- M00-M99 *Diseases of the musculoskeletal system and connective tissue*
- N00-N99 *Diseases of the genitourinary system*
- O00-O9A *Pregnancy, childbirth and the puerperium*
- P00-P96 *Certain conditions originating in the perinatal period*
- Q00-Q99 *Congenital malformations, deformations and chromosomal abnormalities*
- R00-R99 *Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified*
- S00-T88 *Injury, poisoning and certain other consequences of external causes*
- U00-U85 *Codes for special purposes*
- V00-Y99 *External causes of morbidity*
- Z00-Z99 *Factors influencing health status and contact with health services*

# BIBLIOGRAPHIE


Julia Adler-Milstein, A Jay Holmgren, Peter Kralovec, Chantal Worzala, Talisha Searcy, and Vaishali Patel. 2017. Electronic health record adoption in US hospitals: the emergence of a digital "advanced use" divide. *Journal of the American Medical Informatics Association*, 24(6) :1142–1148.

Majid Afshar, Dmitriy Dligach, Brihat Sharma, Xiaoyuan Cai, Jason Boyda, Steven Birch, Daniel Valdez, Suzan Zelisko, Cara Joyce, François Modave, et al. 2019a. Development and application of a high throughput natural language processing architecture to convert all clinical documents in a clinical data warehouse into standardized medical vocabularies. *Journal of the American Medical Informatics Association*, 26(11) :1364–1369.

Majid Afshar, Cara Joyce, Dmitriy Dligach, Brihat Sharma, Robert Kania, Meng Xie, Kristin Swope, Elizabeth Salisbury-Afshar, and Niranjan S Karnik. 2019b. Subtypes in patients with opioid misuse : A prognostic enrichment strategy using electronic health record data in hospitalized patients. *PLoS One*, 14(7) :e0219717.

Naveed Afzal, Vishnu Priya Mallipeddi, Sunghwan Sohn, Hongfang Liu, Rajeev Chaudhry, Christopher G Scott, Iftikhar J Kullo, and Adelaide M Arruda-Olson. 2018. Natural language processing of clinical notes for identification of critical limb ischemia. *International journal of medical informatics*, 111 :83–89.

Naveed Afzal, Sunghwan Sohn, Sara Abram, Christopher G Scott, Rajeev Chaudhry, Hongfang Liu, Iftikhar J Kullo, and Adelaide M Arruda-Olson. 2017. Mining peripheral arterial disease cases from narrative clinical notes using natural language processing. *Journal of vascular surgery*, 65(6) :1753–1761.

Vibhu Agarwal, Tanya Podchiyska, Juan M Banda, Veena Goel, Tiffany I Leung, Evan P Minty, Timothy E Sweeney, Elsie Gyang, and Nigam H Shah. 2016. Learning statistical models of phenotypes using noisy labeled training data. *Journal of the American Medical Informatics Association*, 23(6) :1166–1173.

A Ahmed, C Thongprayoon, BW Pickering, A Akhoundi, G Wilson, D Pieczkiewicz, and V Herasevich. 2014. Towards prevention of acute syndromes. *Applied clinical informatics*, 5(01) :58–72.

Patrick R Alba, Anthony Gao, Kyung Min Lee, Tori Anglin-Foote, Brian Robison, Evangelia Katsoulakis, Brent S Rose, Olga Efimova, Jeffrey P Ferraro, Olga V Patterson, et al. 2021. Ascertainment of veterans with metastatic prostate cancer in electronic health records : demonstrating the case for natural language processing. *JCO clinical cancer informatics*, 5 :1005–1014.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Marie Ansoborlo, Thibault Dhalluin, Christophe Gaborit, Marc Cuggia, and Leslie

Grammatico-Guillon. 2021. Prescreening in oncology using data sciences : The prescious study. In *MIE*, pages 123–127.

Wissam Antoun, Benoît Sagot, and Djamé Seddah. 2023a. Data-efficient French language modeling with CamemBERTa. In *Findings of the Association for Computational Linguistics : ACL 2023*, pages 5174–5185, Toronto, Canada. Association for Computational Linguistics.

Wissam Antoun, Benoît Sagot, and Djamé Seddah. 2023b. Data-Efficient French Language Modeling with CamemBERTa. In *Findings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL'23)*, Toronto, Canada.

Mattia Arrigo, Mariell Jessup, Wilfried Mullens, Nosheen Reza, Ajay M Shah, Karen Sliwa, and Alexandre Mebazaa. 2020. Acute heart failure. *Nature Reviews Disease Primers*, 6(1) :16.

Naveen Ashish, Lisa Dahm, and Charles Boicey. 2014. University of california, irvine–pathology extraction pipeline : The pathology extraction pipeline for information extraction from pathology reports. *Health informatics journal*, 20(4) :288–305.

Ye Seul Bae, Kyung Hwan Kim, Han Kyul Kim, Sae Won Choi, Taehoon Ko, Hee Hwa Seo, Hae-Young Lee, and Hyojin Jeon. 2021. Keyword extraction algorithm for classifying smoking status from unstructured bilingual electronic health records based on natural language processing. *Applied Sciences*, 11(19) :8812.

Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan H"ogberg, Ulla Stenius, and Anna Korhonen. 2016. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinform.*, 32(3) :432–440.

Lisa Bastarache, Jacob J Hughey, Jeffrey A Goldstein, Julie A Bastraache, Satya Das, Neil Charles Zaki, Chenjie Zeng, Leigh Anne Tang, Dan M Roden, and Joshua C Denny. 2019. Improving the phenotype risk score as a scalable approach to identifying patients with mendelian disease. *Journal of the American Medical Informatics Association*, 26(12) :1437–1447.

Adrien Bazoge. 2021. Revue de la littérature : entrepôts de données biomédicales et traitement automatique de la langue (literature review : biomedical data warehouse and natural language processing ). In *Actes de la 28e Conférence sur le Traitement*

*Automatique des Langues Naturelles. Volume 2 : 23e REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL)*, pages 96–109, Lille, France. ATALA.

Adrien Bazoge, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. Natural language processing on data from clinical data warehouse : A systematic review. *JMIR Medical Informatics*, -(-) :to appear.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert : A pretrained language model for scientific text. *arXiv preprint arXiv :1903.10676*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer : The long-document transformer. *arXiv :2004.05150*.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null) :1137–1155.

Aman Berhe, Guillaume Draznieks, Vincent Martenot, Valentin Masdeu, Lucas Davy, and Jean-Daniel Zucker. 2023. AliBERT: A pre-trained language model for French biomedical text. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 223–236, Toronto, Canada. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls) : integrating biomedical terminology. *Nucleic acids research*, 32 Database issue :D267–70.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5 :135–146.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar

Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the opportunities and risks of foundation models.

Robert Bossy, Louise Deléger, Estelle Chaix, Mouhamadou Ba, and Claire Nédellec. 2019. Bacteria biotope at BioNLP open shared tasks 2019. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 121–131, Hong Kong, China. Association for Computational Linguistics.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Selen Bozkurt, Kathleen M Kan, Michelle K Ferrari, Daniel L Rubin, Douglas W Blayney, Tina Hernandez-Boussard, and James D Brooks. 2019. Is it possible to automatically assess pretreatment digital rectal examination documentation using natural language processing ? a single-centre retrospective study. *BMJ open*, 9(7) :e027182.

Selen Bozkurt, Rohan Paul, Jean Coquet, Ran Sun, Imon Banerjee, James D Brooks, and Tina Hernandez-Boussard. 2020. Phenotyping severity of patient-centered outcomes using clinical notes : A prostate cancer use case. *Learning Health Systems*, 4(4) :e10237.

Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. 2015a. Extraction of relations between genes and diseases from text and large-scale data analysis : implications for translational research. *BMC bioinformatics*, 16 :1–17.

Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. 2015b. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, 16(1).

Keno K. Bressem, Jens-Michalis Papaioannou, Paul Grundmann, Florian Borchert, Lisa C. Adams, Leonhard Liu, Felix Busch, Lina Xu, Jan P. Loyen, Stefan M. Niehues, Moritz Augustin, Lennart Grosser, Marcus R. Makowski, Hugo JWL. Aerts, and Alexander Löser. 2023. Medbert.de: A comprehensive german bert model for the medical domain. *arXiv preprint arXiv :2303.08179*. Keno K. Bressem and Jens-Michalis Papaioannou and Paul Grundmann contributed equally.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Andrea Caccamisi, Leif Jørgensen, Hercules Dalianis, and Mats Rosenlund. 2020. Natural language processing and machine learning to enable automatic extraction and classification of patients' smoking status from electronic medical records. *Upsala Journal of Medical Sciences*, 125(4) :316–324. PMID : 32696698.

Deniz Caliskan, Jakob Zierk, Detlef Kraska, Stefan Schulz, Philipp Daumke, Hans-Ulrich Prokosch, and Lorenz A Kapsner. 2021. First steps to evaluate an nlp tool's medication extraction accuracy from discharge letters. *Stud Health Technol Inform*, 278 :224–230.

Boris Campillo-Gimenez, Nicolas Garcelon, Pascal Jarno, Jean Marc Chapplain, and Marc Cuggia. 2012. Full-text automated detection of surgical site infections secondary to neurosurgery in rennes, france. *Studies in health technology and informatics*, 192 :572–5.

Hui Cao, Marianthi Markatou, Genevieve B Melton, Michael F Chiang, and George Hripcsak. 2005. Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. In *AMIA Annual Symposium Proceedings*, volume 2005, page 106. American Medical Informatics Association.

Rémi Cardon, Natalia Grabar, Cyril Grouin, and Thierry Hamon. 2020. Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques (presentation of the DEFT 2020 challenge : open domain textual similarity and precise information extraction from clinical cases ). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, pages 1–13, Nancy, France. ATALA et AFCP.

Lorena Carlo, Herbert S Chase, and Chunhua Weng. 2010. Aligning structured and unstructured medical problems using umls. In *AMIA Annual Symposium Proceedings*, volume 2010, page 91. American Medical Informatics Association.

Casimiro Pio Carrino, Jordi Armengol-Estapé, Asier Gutiérrez-Fandiño, Joan Llop-Palao, Marc Pàmies, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Biomedical and clinical language models for spanish : On the benefits of domain-specific pretraining in a mid-resource scenario. *arXiv preprint arXiv :2109.03570*.

Gebra Cuyun Carter, Pamela B Landsman-Blumberg, Barbara H Johnson, Paul Juneau, Steven J Nicol, Li Li, and Veena Shankaran. 2015. Kras testing of patients with metastatic colorectal cancer in a community-based oncology setting : a retrospective database analysis. *Journal of Experimental & Clinical Cancer Research*, 34(1) :1–8.

Anne B Casto. 2013. *Principles of healthcare reimbursement*. Citeseer.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert : The muppets straight out of law school. *arXiv preprint arXiv :2010.02559*.

Alec B Chapman, Audrey Jones, A Taylor Kelley, Barbara Jones, Lori Gawron, Ann Elizabeth Montgomery, Thomas Byrne, Ying Suo, James Cook, Warren Pettey, Kelly Peterson, Makoto Jones, and Richard Nelson. 2021. Rehoused: A novel measurement of veteran housing stability using natural language processing. *Journal of Biomedical Informatics*, 122 :103903.

Herbert S Chase, Lindsey R Mitrani, Gabriel G Lu, and Dominick J Fulgieri. 2017. Early

recognition of multiple sclerosis using natural language processing of the electronic health record. *BMC medical informatics and decision making*, 17(1) :1–8.

Herbert S Chase, Jai Radhakrishnan, Shayan Shirazian, Maya K Rao, and David K Vawdrey. 2010. Under-documentation of chronic kidney disease in the electronic health record in outpatients. *Journal of the American Medical Informatics Association*, 17(5) :588–594.

Chi-Jen Chen, Neha Warikoo, Yung-Chun Chang, Jin-Hua Chen, and Wen-Lian Hsu. 2019a. Medical knowledge infused convolutional neural networks for cohort selection in clinical trials. *Journal of the American Medical Informatics Association*, 26(11) :1227–1236.

Elizabeth S Chen, George Hripcsak, Hua Xu, Marianthi Markatou, and Carol Friedman. 2008. Automated acquisition of disease–drug knowledge from biomedical and clinical documents : an initial study. *Journal of the American Medical Informatics Association*, 15(1) :87–98.

Jonathan H Chen, Mary K Goldstein, Steven M Asch, Lester Mackey, and Russ B Altman. 2017. Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets. *Journal of the American Medical Informatics Association*, 24(3) :472–480.

Long Chen, Yu Gu, Xin Ji, Chao Lou, Zhiyong Sun, Haodan Li, Yuan Gao, and Yang Huang. 2019b. Clinical trial cohort selection based on multi-level rule-based natural language processing system. *Journal of the American Medical Informatics Association*, 26(11) :1218–1226.

Q. Chen, A. Allot, and Z. Lu. 2020a. Keep up with the latest coronavirus research. *Nature*, 579(7798) :193.

Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, Meng Zhou, Penghui Zhu, and Pengtao Xie. 2020b. Meddialog: A large-scale medical dialogue dataset. *CoRR*, abs/2004.03329.

Wansu Chen, Rebecca K Butler, Yichen Zhou, Rex A Parker, Christie Y Jeon, and Bechien U Wu. 2020c. Prediction of pancreatic cancer based on imaging features in patients with duct abnormalities. *Pancreas*, 49(3) :413.

Xiaoyi Chen, Nicolas Garcelon, Antoine Neuraz, Katy Billot, Marc Lelarge, Thomas Bonald, Hugo Garcia, Yoann Martin, Vincent Benoit, Marc Vincent, et al. 2019c. Phenotypic similarity for rare disease : ciliopathy diagnoses and subtyping. *Journal of Biomedical Informatics*, 100 :103308.

Raj Chetty, Michael Stepner, Sarah Abraham, Shelby Lin, Benjamin Scuderi, Nicholas Turner, Augustin Bergeron, and David Cutler. 2016. The Association Between Income and Life Expectancy in the United States, 2001-2014. *JAMA*, 315(16) :1750–1766.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv :1904.10509*.

Billy Chiu, Sampo Pyysalo, Ivan Vulić, and Anna Korhonen. 2018. Bio-simverb and bio-simlex : wide-coverage evaluation sets of word similarity in biomedicine. *BMC bioinformatics*, 19(1) :1–13.

Falgun H Chokshi, Bonggun Shin, Timothy Lee, Andrew Lemmon, Sean Necessary, and Jinho D Choi. 2017. Natural language processing for classification of acute, communicable findings on unstructured head ct reports : comparison of neural network and non-neural machine learning techniques. *bioRxiv*, page 173310.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. 2020. Rethinking attention with performers. *arXiv preprint arXiv :2009.14794*.

Laurent Chouchana, Nathanaël Beeker, Nicolas Garcelon, Bastien Rance, Nicolas Paris, Elisa Salamanca, Elisabeth Polard, Anita Burgun, Jean-Marc Treluyer, Antoine Neuraz, et al. 2021. Association of antihypertensive agents with the risk of in-hospital death in patients with covid-19. *Cardiovascular drugs and therapy*, pages 1–6.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm : Scaling language modeling with pathways. *arXiv preprint arXiv :2204.02311*.

Jan Chrusciel, François Girardon, Lucien Roquette, David Laplanche, Antoine Duclos, and Stéphane Sanchez. 2021. The prediction of hospital length of stay using unstructured data. *BMC Medical Informatics and Decision Making*, 21(1) :351.

Jen-Hsiang Chuang, Carol Friedman, and George Hripcsak. 2002. A comparison of the charlson comorbidities derived from medical language processing and administrative data. In *Proceedings of the AMIA Symposium*, page 160. American Medical Informatics Association.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Kenneth Ward Church. 2020. Emerging trends: Subwords, seriously? *Natural Language Engineering*, 26(3) :375–382.

Marta Luisa Ciofi Degli Atti, Fabrizio Pecoraro, Simone Piga, Daniela Luzi, and Massimiliano Raponi. 2020. Developing a surgical site infection surveillance system based on hospital unstructured clinical notes and text mining. *Surgical Infections*, 21(8) :716–721.

Laila R Cochon, Catherine S Giess, and Ramin Khorasani. 2020. Comparing diagnostic performance of digital breast tomosynthesis and full-field digital mammography. *Journal of the American College of Radiology*, 17(8) :999–1003.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1) :37–46.

Raphael Cohen, Michael Elhadad, and Noémie Elhadad. 2013. Redundancy in electronic health record corpora : analysis, impact on text mining performance and mitigation strategies. *BMC bioinformatics*, 14(1) :1–15.

Nigel Collier and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78.

Commission nationale informatique et des libertés. 2021. Référentiel relatif aux traitements de données à caractère personnel mis en oeuvre à des fins de création d'entrepôts de données dans le domaine de la santé. `https://www.cnil.fr/sites/default/files/atoms/files/referentiel_entrepot.pdf`.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jean Coquet, Selen Bozkurt, Kathleen M Kan, Michelle K Ferrari, Douglas W Blayney, James D Brooks, and Tina Hernandez-Boussard. 2019. Comparison of orthogonal nlp methods for clinical phenotyping and assessment of bone scan utilization among prostate cancer patients. *Journal of biomedical informatics*, 94 :103184.

Sébastien Cossin, Margaux Jolly, and Iban Larrouture. 2021. Semi-automatic extraction of abbreviations and their senses from electronic health records. *preprint*.

Noa P. Cruz Díaz and Manuel Maña López. 2015. An analysis of biomedical tokenization: Problems and strategies. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 40–49, Lisbon, Portugal. Association for Computational Linguistics.

Clément Dalloux, Vincent Claveau, Natalia Grabar, Lucas Emanuel Silva Oliveira, Claudia Maria Cabral Moro, Yohan Bonescki Gumiel, and Deborah Ribeiro Carvalho. 2021. 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.

Jessica K De Freitas, Kipp W Johnson, Eddye Golden, Girish N Nadkarni, Joel T Dudley, Erwin P Bottinger, Benjamin S Glicksberg, and Riccardo Miotto. 2021. Phe2vec : Automated disease phenotyping based on unsupervised embeddings from electronic health records. *Patterns*, 2(9) :100337.

Dina Demner-Fushman, Sameer Antani, Matthew Simpson, and George R Thoma. 2012. Design and development of a multimodal biomedical information retrieval system. *Journal of Computing Science and Engineering*, 6(2) :168–177.

Henri Denolin, Horst Kuhn, H Krayenbuehl, Franz Loogen, and Attilio Reale. 1983. The defintion of heart failure. *European heart journal*, 4(7) :445–448.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings*

*of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132, Online. Association for Computational Linguistics.

Dmitriy Dligach, Majid Afshar, and Timothy Miller. 2019. Toward a clinical text encoder : pretraining for clinical natural language processing with applications to substance misuse. *Journal of the American Medical Informatics Association*, 26(11) :1272–1278.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus : a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47 :1–10.

Basile Dura, Charline Jean, Xavier Tannier, Alice Calliger, Romain Bey, Antoine Neuraz, and Rémi Flicoteaux. 2022. Learning structures of the french clinical language:development and validation of word embedding models using 21 million clinical reports from electronic health records.

Jean-Charles Duthe, Guillaume Bouzille, Emmanuelle Sylvestre, Emmanuel Chazard, Cedric Arvieux, and Marc Cuggia. 2021. How to identify potential candidates for hiv pre-exposure prophylaxis : an ai algorithm reusing real-world hospital data. *Studies in Health Technology and Informatics*, 281 :714–718.

Hicham El Boukkouri. 2021. *Domain adaptation of word embeddings through the exploitation of in-domain corpora and knowledge bases*. Theses, Université Paris-Saclay.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, and Pierre Zweigenbaum. 2022. Re-train or train from scratch? comparing pre-training strategies of BERT in the medical domain. In *Proceedings of the Thirteenth Language Resources and Evaluation*

*Conference*, pages 2626–2633, Marseille, France. European Language Resources Association.

Inc. Elucidata. 2020. Geokhoj v1. `https://github.com/ElucidataInc/GEOKhoj-datasets/tree/main/geokhoj_v1`.

Jean-Baptiste Escudié, Anne-Sophie Jannot, Eric Zapletal, Sarah Cohen, Georgia Malamut, Anita Burgun, and Bastien Rance. 2015. Reviewing 741 patients records in two hours with fastvisu. In *AMIA Annual Symposium Proceedings*, volume 2015, page 553. American Medical Informatics Association.

R Scott Evans, Jose Benuzillo, Benjamin D Horne, James F Lloyd, Alejandra Bradshaw, Deborah Budge, Kismet D Rasmusson, Colleen Roberts, Jason Buckway, Norma Geer, et al. 2016. Automated identification and predictive tools to help identify high-risk heart failure patients : pilot evaluation. *Journal of the American Medical Informatics Association*, 23(5) :872–878.

Sabri Eyuboglu, Geoffrey Angus, Bhavik N Patel, Anuj Pareek, Guido Davidzon, Jin Long, Jared Dunnmon, and Matthew P Lungren. 2021. Multi-task weak supervision enables anatomically-resolved abnormality detection in whole-body fdg-pet/ct. *Nature communications*, 12(1) :1880.

Martin Faltys, Marc Zimmermann, Xinrui Lyu, Matthias Hüser, Stephanie Hyland, Gunnar Rätsch, and Tobias Merz. 2021. Hirid, a high time-resolution icu dataset (version 1.1. 1). *PhysioNet*.

Alvan R Feinstein and Domenic V Cicchetti. 1990. High agreement but low kappa : I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6) :543–549.

Daniel J Feller, Jason Zucker, Bharat Srikishan, Roxana Martinez, Henry Evans, Michael T Yin, Peter Gordon, Noémie Elhadad, et al. 2018a. Towards the inference of social and behavioral determinants of sexual health : development of a gold-standard corpus with semi-supervised learning. In *AMIA Annual Symposium Proceedings*, volume 2018, page 422. American Medical Informatics Association.

Daniel J Feller, Jason Zucker, Michael T Yin, Peter Gordon, and Noémie Elhadad. 2018b. Using clinical notes and natural language processing for automated hiv risk assessment. *Journal of acquired immune deficiency syndromes (1999)*, 77(2) :160.

Daniel J. Feller, Jason E. Zucker, Oliver Bear Don't Walk, Bharat Srikishan, Roxana Martinez, Henry Evans, Michael T. Yin, Peter Gordon, and Noémie Elhadad. 2018c. Towards the inference of social and behavioral determinants of sexual health : Development of a gold-standard corpus with semi-supervised learning. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2018 :422–429.

Oscar Ferrández, Brett R South, Shuying Shen, F Jeffrey Friedlin, Matthew H Samore, and Stéphane M Meystre. 2013. Bob, a best-of-breed automated text de-identification system for vha clinical documents. *Journal of the American Medical Informatics Association*, 20(1) :77–83.

Thomas Ferté, Sébastien Cossin, Thierry Schaeverbeke, Thomas Barnetche, Vianney Jouhet, and Boris P Hejblum. 2021. Automatic phenotyping of electronical health record : Phevis algorithm. *Journal of Biomedical Informatics*, 117 :103746.

Johanna Fiebeck, Hans Laser, Hinrich B Winther, and Svetlana Gerbel. 2018. Leaving no stone unturned : using machine learning based approaches for information extraction from full texts of a research data warehouse. In *International Conference on Data Integration in the Life Sciences*, pages 50–58. Springer.

Rosa L. Figueroa, Diego A. Soto, and Esteban J. Pino. 2014. Identifying and extracting patient smoking status information from clinical narrative texts in spanish. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2710–2713.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5) :378.

Jason Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Sunny Kang, Rosaline Su, Wojciech Kusa, Samuel Cahyawijaya, et al. 2022. Bigbio : a framework for data-centric biomedical natural language processing. *Advances in Neural Information Processing Systems*, 35 :25792–25806.

Kristina K Gagalova, M Angelica Leon Elizalde, Elodie Portales-Casamar, and Matthias Görges. 2020a. What you need to know before implementing a clinical research data warehouse : comparative review of integrated data repositories in health care institutions. *JMIR formative research*, 4(8) :e17687.

Kristina K Gagalova, M Angelica Leon Elizalde, Elodie Portales-Casamar, and Matthias Görges. 2020b. What you need to know before implementing a clinical research data warehouse : Comparative review of integrated data repositories in health care institutions. *JMIR Form. Res.*, 4(8) :e17687.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2) :23–38.

Nicolas Garcelon, Antoine Neuraz, Vincent Benoit, Rémi Salomon, and Anita Burgun. 2017a. Improving a full-text search engine : the importance of negation detection and family history context to identify cases in a biomedical data warehouse. *Journal of the American Medical Informatics Association*, 24(3) :607–613.

Nicolas Garcelon, Antoine Neuraz, Vincent Benoit, Rémi Salomon, Sven Kracker, Felipe Suarez, Nadia Bahi-Buisson, Smail Hadj-Rabia, Alain Fischer, Arnold Munnich, et al. 2017b. Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse : Dr. warehouse and the needle in the needle stack. *Journal of biomedical informatics*, 73 :51–61.

Nicolas Garcelon, Antoine Neuraz, Rémi Salomon, Nadia Bahi-Buisson, Jeanne Amiel, Capucine Picard, Nizar Mahlaoui, Vincent Benoit, Anita Burgun, and Bastien Rance. 2018. Next generation phenotyping using narrative reports in a rare disease clinical data warehouse. *Orphanet journal of rare diseases*, 13(1) :1–11.

Sebastian Gehrmann, Franck Dernoncourt, Yeran Li, Eric T. Carlson, Joy T. Wu, Jonathan Welt, John Foote, Jr., Edward T. Moseley, David W. Grant, Patrick D. Tyler, and Leo A. Celi. 2018. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PLOS ONE*, 13(2) :1–19.

N Genes, D Chandra, S Ellis, and K Baumlin. 2013. Validating emergency department vital signs using a data quality engine for data warehouse. *The open medical informatics journal*, 7 :34.

Martin Gerner, Goran Nenadic, and Casey M Bergman. 2010. Linnaeus : a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1) :1–17.

Alon Geva, Steven H Abman, Shannon F Manzi, Dunbar D Ivy, Mary P Mullen, John Griffin, Chen Lin, Guergana K Savova, and Kenneth D Mandl. 2020. Adverse drug event rates in pediatric pulmonary hypertension : a comparison of real-world data sources. *Journal of the American Medical Informatics Association*, 27(2) :294–300.

Alexander P Glaser, Brian J Jordan, Jason Cohen, Anuj Desai, Philip Silberman, and Joshua J Meeks. 2018. Automated extraction of grade, stage, and quality information from transurethral resection of bladder tumor pathology reports using natural language processing. *JCO Clinical Cancer Informatics*, 2 :1–8.

Sara Bersche Golas, Takuma Shibahara, Stephen Agboola, Hiroko Otaki, Jumpei Sato, Tatsuya Nakae, Toru Hisamitsu, Go Kojima, Jennifer Felsted, Sujay Kakarmath, et al. 2018. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure : a retrospective analysis of electronic medical records data. *BMC medical informatics and decision making*, 18(1) :1–17.

Sigfried Gold, Noémie Elhadad, Xinxin Zhu, James J Cimino, and George Hripcsak. 2008. Extracting structured medication event information from discharge summaries. In *AMIA Annual Symposium Proceedings*, volume 2008, page 237. American Medical Informatics Association.

Natalia Grabar, Vincent Claveau, and Clément Dalloux. 2018. CAS: French Corpus with Clinical Cases. In *Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 1–7, Brussels, Belgium.

Natalia Grabar, Cyril Grouin, Thierry Hamon, and Vincent Claveau. 2019. Recherche et extraction d'information dans des cas cliniques. présentation de la campagne d'évaluation deft 2019. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Défi Fouille de Textes (atelier TALN-RECITAL)*, pages 7–16, Toulouse, France. Association pour le Traitement Automatique des Langues. Information Retrieval and Information Extraction from Clinical Cases.

Cyril Grouin, Natalia Grabar, and Gabriel Illouz. 2021. Classification de cas cliniques et évaluation automatique de réponses d'étudiants : présentation de la campagne DEFT 2021 (clinical cases classification and automatic evaluation of student answers : Presentation of the DEFT 2021 challenge). In *Actes de la 28e Conférence sur le Traitement*

*Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT)*, pages 1–13, Lille, France. ATALA.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Morgan Guillaudeux, Olivia Rousseau, Julien Petot, Zineb Bennis, Charles-Axel Dein, Thomas Goronflot, Nicolas Vince, Sophie Limou, Matilde Karakachoff, Matthieu Wargny, et al. 2023. Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. *NPJ Digital Medicine*, 6(1) :37.

Tracy D Gunter and Nicolas P Terry. 2005. The emergence of national electronic health record architectures in the united states and australia: Models, costs, and questions. *J Med Internet Res*, 7(1) :e3.

Kilem Gwet. 2001. Handbook of inter-rater reliability : How to estimate the level of agreement between two or multiple raters. *Gaithersburg, MD : STATAXIS Publishing Company.*

Kilem L Gwet. 2014. *Handbook of inter-rater reliability : The definitive guide to measuring the extent of agreement among raters.* Advanced Analytics, LLC.

Birger Haarbrandt, Erik Tute, and Michael Marschollek. 2016. Automated population of an i2b2 clinical data warehouse from an openehr-based data repository. *Journal of Biomedical Informatics*, 63 :277–294.

Krystl Haerian, Hojjat Salmasian, and Carol Friedman. 2012. Methods for identifying suicide or suicidal ideation in ehrs. In *AMIA annual symposium proceedings*, volume 2012, page 1244. American Medical Informatics Association.

Alaa Hamoud, Ali Salah Hashim, and Wid Akeel Awadh. 2018. Clinical data warehouse : a review. *Iraqi Journal for Computers and Informatics*, 44(2).

Sifei Han, Robert F. Zhang, Lingyun Shi, Russell Richie, Haixia Liu, Andrew Tseng, Wei Quan, Neal Ryan, David Brent, and Fuchiang R. Tsui. 2022. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. *Journal of Biomedical Informatics*, 127 :103984.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2023. Medalpaca – an open-source collection of medical conversational ai models and training data.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, China. Association for Computational Linguistics.

Rave Harpaz, Santiago Vilar, William DuMouchel, Hojjat Salmasian, Krystl Haerian, Nigam H Shah, Herbert S Chase, and Carol Friedman. 2013. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. *Journal of the American Medical Informatics Association*, 20(3) :413–419.

Daniel R Harris, Darren W Henderson, and Alexandria Corbeau. 2020. sig2db : a workflow for processing natural language from prescription instructions for clinical data warehouses. *AMIA Summits on Translational Science Proceedings*, 2020 :221.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3) :146–162.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Tiancheng He, Joy Nolte Fong, Linda W Moore, Chika F Ezeana, David Victor, Mukul Divatia, Matthew Vasquez, R Mark Ghobrial, and Stephen TC Wong. 2021b. An imageomics and multi-network based deep learning model for risk assessment of liver transplantation for hepatocellular cancer. *Computerized Medical Imaging and Graphics*, 89 :101894.

Tiancheng He, Mamta Puppala, Chika F Ezeana, Yan-siang Huang, Ping-hsuan Chou, Xiaohui Yu, Shenyi Chen, Lin Wang, Zheng Yin, Rebecca L Danforth, et al. 2019. A deep learning–based decision support tool for precision risk assessment of breast cancer. *JCO clinical cancer informatics*, 3 :1–12.

Tiancheng He, Mamta Puppala, Richard Ogunti, James J Mancuso, Xiaohui Yu, Shenyi Chen, Jenny C Chang, Tejal A Patel, and Stephen TC Wong. 2017. Deep learning analytics for diagnostic support of breast cancer disease management. In *2017 IEEE EMBS international conference on biomedical & health informatics (BHI)*, pages 365–368. IEEE.

Serge Heiden, Jean-Philippe Magué, and Bénédicte Pincemin. 2010. TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. In *10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*, volume 2, pages 1021–1032, Rome, Italy. Edizioni Universitarie di Lettere Economia Diritto.

William Hersh, Chris Buckley, T. J. Leone, and David Hickam. 1994. Ohsumed : An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, page 192–201, Berlin, Heidelberg. Springer-Verlag.

Nicolas Hiebel, Olivier Ferret, Karën Fort, and Aurélie Névéol. 2022. CLISTER : A corpus for semantic textual similarity in French clinical narratives. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4306–4315, Marseille, France. European Language Resources Association.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8) :1735–1780.

Jet Hoek and Merel Scholman. 2017. Evaluating discourse annotation : Some recent insights and new approaches. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (isa-13)*.

Nicolas Hoertel, Marina Sánchez-Rico, Raphaël Vernet, Nathanaël Beeker, Antoine Neuraz, Jesús M Alvarado, Christel Daniel, Nicolas Paris, Alexandre Gramfort, Guillaume Lemaitre, et al. 2021. Dexamethasone use and mortality in hospitalized patients with coronavirus disease 2019 : A multicentre retrospective observational study. *British journal of clinical pharmacology*, 87(10) :3766–3775.

Sarah R Hoffman, Anissa I Vines, Jacqueline R Halladay, Emily Pfaff, Lauren Schiff, Daniel Westreich, Aditi Sundaresan, La-Shell Johnson, and Wanda K Nicholson. 2018. Optimizing research in symptomatic uterine fibroids with development of a computable

phenotype for use with electronic health records. *American journal of obstetrics and gynecology*, 218(6) :610–e1.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv :2203.15556*.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 3594–3608, Online. Association for Computational Linguistics.

John H Holmes, Thomas E Elliott, Jeffrey S Brown, Marsha A Raebel, Arthur Davidson, Andrew F Nelson, Annie Chung, Pierre La Chance, and John F Steiner. 2014. Clinical research data warehouse governance for distributed research networks in the usa : a systematic review of the literature. *Journal of the American Medical Informatics Association*, 21(4) :730–736.

Sung Noh Hong, Hee Jung Son, Sun Kyu Choi, Dong Kyung Chang, Young-Ho Kim, Sin-Ho Jung, and Poong-Lyul Rhee. 2017. A prediction model for advanced colorectal neoplasia in an asymptomatic screening population. *PloS one*, 12(8) :e0181040.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3) :296–298.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert : Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv :1904.05342*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. Clinicalbert: Modeling clinical notes and predicting hospital readmission.

Marie Hully, Tommaso Lo Barco, Anna Kaminska, Giulia Barcia, Claude Cances, Cyril Mignot, Isabelle Desguerre, Nicolas Garcelon, Edor Kabashi, and Rima Nabbout. 2021. Deep phenotyping unstructured data mining in an extensive pediatric database to unravel a common kcna2 variant in neurodevelopmental syndromes. *Genetics in Medicine*, 23(5) :968–971.

Haley S Hunter-Zinck, Jordan S Peck, Tania D Strout, and Stephan A Gaehde. 2019. Predicting emergency department orders with multilabel machine learning techniques and simulating effects on length of stay. *Journal of the American Medical Informatics Association*, 26(12) :1427–1436.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have ? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14) :6421.

A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark. 2020. MIMIC-IV (version 0.3). `https://doi.org/10.13026/a2mm-bn44`.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Sci. Data*, 3(1) :160035.

Siddhartha Jonnalagadda, Trevor Cohen, Stephen Wu, and Graciela Gonzalez. 2012. Enhancing clinical concept extraction with distributional semantics. *Journal of biomedical informatics*, 45(1) :129–140.

Siddhartha Jonnalagadda, Trevor Cohen, Stephen Wu, Hongfang Liu, and Graciela Gonzalez. 2013. Using empirically constructed lexical resources for named entity recognition. *Biomedical informatics insights*, 6 :BII–S11664.

Jordan Jouffroy, Sarah F Feldman, Ivan Lerner, Bastien Rance, Anita Burgun, Antoine Neuraz, et al. 2021. Hybrid deep learning for medication-related information extraction from clinical texts in french : Medext algorithm development study. *JMIR medical informatics*, 9(3) :e17934.

Young Juhn and Hongfang Liu. 2020. Artificial intelligence approaches using natural language processing to advance ehr-based clinical research. *Journal of Allergy and Clinical Immunology*, 145(2) :463–469.

Kenneth Jung, Paea LePendu, Srinivasan Iyer, Anna Bauer-Mehren, Bethany Percha, and Nigam H Shah. 2015. Functional evaluation of out-of-the-box text-mining tools for data-mining tasks. *Journal of the American Medical Informatics Association*, 22(1) :121–131.

Kenneth Jung, Paea LePendu, and Nigam Shah. 2013. Automated detection of systematic off-label drug use in free text of electronic medical records. *AMIA Summits on Translational Science Proceedings*, 2013 :94.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv :2001.08361*.

A. Karlsson, A. Ellonen, H. Irjala, V. Väliaho, K. Mattila, L. Nissi, E. Kytö, S. Kurki, R. Ristamäki, P. Vihinen, T. Laitinen, A. Ålgars, S. Jyrkkiö, H. Minn, and E. Heervä. 2021. Impact of deep learning-determined smoking status on mortality of cancer patients: never too late to quit. *ESMO Open*, 6(3) :100175.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are rnns : Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail : A textual entailment dataset from science question answering. In *AAAI*.

Sunder Ali Khowaja, Parus Khuwaja, and Kapal Dev. 2023. Chatgpt needs spade (sustainability, privacy, digital divide, and ethics) evaluation: A review.

Chulho Kim, Vivienne Zhu, Jihad Obeid, and Leslie Lenert. 2019a. Natural language processing and machine learning algorithm to identify brain mri reports with acute ischemic stroke. *PloS one*, 14(2) :e0212778.

Ellen Kim, Samuel M Rubinstein, Kevin T Nead, Andrzej P Wojcieszynski, Peter E Gabriel, and Jeremy L Warner. 2019b. The evolving use of electronic health records (ehr) for research. *Seminars in radiation oncology*, 29(4) :354–361.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1) :i180–i182.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer : The efficient transformer. *arXiv preprint arXiv :2001.04451*.

Eyal Klang, Benjamin R Kummer, Neha S Dangayach, Amy Zhong, M Arash Kia, Prem Timsina, Ian Cossentino, Anthony B Costa, Matthew A Levin, and Eric K Oermann.

2021. Predicting adult neuroscience intensive care unit admission from emergency department triage using a retrospective, tabular-free text machine learning approach. *Scientific reports*, 11(1) :1381.

Eva S Klappe, Florentien JP van Putten, Nicolette F de Keizer, and Ronald Cornet. 2021. Contextual property detection in dutch diagnosis descriptions for uncertainty, laterality and temporality. *BMC Medical Informatics and Decision Making*, 21(1) :1–17.

Alican Kocabiyikoglu, François Portet, Prudence Gibert, Hervé Blanchon, Jean-Marc Babouchkine, and Gaëtan Gavazzi. 2022. A spoken drug prescription dataset in french for spoken language understanding. In *13th Language Resources and Evaluation Conference (LREC 2022)*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Felix Köpcke and Hans-Ulrich Prokosch. 2014. Employing computers for the recruitment into clinical trials : a comprehensive systematic review. *Journal of medical Internet research*, 16(7) :e161.

Jan A Kors, Simon Clematide, Saber A Akhondi, Erik M van Mulligen, and Dietrich Rebholz-Schuhmann. 2015. A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. *Journal of the American Medical Informatics Association*, 22(5) :948–956.

Martin Krallinger, Obdulia Rabal, Saber A Akhondi, Martın Pérez Pérez, Jesús Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, José Antonio López, Umesh Nandal, et al. 2017. Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, volume 1, pages 141–146.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015.

The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1) :1–17.

Jonathan Krebs, Max Bittrich, Georg Dietrich, Maximilian Ertl, Georg Fette, Mathias Kaspar, Leon Liman, Hermann Einsele, Frank Puppe, and Stefan Knop. 2018. Finding needles in the haystack : Identifying patients with rare subtype of multiple myeloma supported by a data warehouse and information extraction. In *GMDS*, pages 160–164.

Bhavani Singh Agnikula Kshatriya, Joyce E Balls-Berry, William D Freeman, Rui Zhang, and Yanshan Wang. 2020. Completeness of social and behavioral determinants of health in electronic health records : A case study on the patient-provided information from a minority cohort with sexually transmitted diseases. *Preprint*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Béatrice Daille, Pierre-Antoine Gourraud, Emmanuel Morin, and Mickael Rouvier. 2022. FrenchMedMCQA: A French Multiple-Choice Question Answering Dataset for Medical domain. In *LOUHI 2022 @ Empirical Methods in Natural Language Processing (EMNLP) 2022*, Abou Dhabi, United Arab Emirates.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023a. DrBERT: A robust pre-trained model in French for biomedical and clinical domains. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 16207–16221, Toronto, Canada. Association for Computational Linguistics.

Yanis Labrak, Mickaël Rouvier, and Richard Dufour. 2023b. MORFITT : A multi-label corpus of French scientific articles in the biomedical domain. In *30e Conférence sur*

*le Traitement Automatique des Langues Naturelles (TALN) Atelier sur l'Analyse et la Recherche de Textes Scientifiques*, Paris, France. Florian Boudin.

Ronilda Lacson, Aijia Wang, Laila Cochon, Catherine Giess, Sonali Desai, Sunil Eappen, and Ramin Khorasani. 2020. Factors associated with optimal follow-up in women with bi-rads 3 breast findings. *Journal of the American College of Radiology*, 17(4) :469–474.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Jörg Landthaler, Bernhard Waltl, Dominik Huth, Daniel Braun, Christoph Stocker, Thomas Geiger, and Florian Matthes. 2017. Extending thesauri using word embeddings and the intersection method. *ASAIL@ ICAIL*, 8(1) :112–119.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553) :436–444.

Dongha Lee, Xiaoqian Jiang, and Hwanjo Yu. 2020a. Harmonized representation learning on dynamic ehr graphs. *Journal of biomedical informatics*, 106 :103426.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4) :1234–1240.

Kye Hwa Lee, Hyo Jung Kim, Yi-Jun Kim, Ju Han Kim, and Eun Young Song. 2020b. Extracting structured genotype information from free-text hla reports using a rule-based approach. *Journal of Korean Medical Science*, 35(12).

Yeong Chan Lee, Sang-Hyuk Jung, Aman Kumar, Injeong Shim, Minku Song, Min Seo Kim, Kyunga Kim, Woojae Myung, Woong-Yang Park, and Hong-Hee Won. 2023. Icd2vec : Mathematical representation of diseases. *Journal of Biomedical Informatics*, 141 :104361.

Nicholas J Leeper, Anna Bauer-Mehren, Srinivasan V Iyer, Paea LePendu, Cliff Olson, and Nigam H Shah. 2013. Practice-based evidence : profiling the safety of cilostazol by text-mining of clinical notes. *PloS one*, 8(5) :e63499.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do we still need clinical language models? In *Proceedings of the Conference on Health, Inference, and Learning*, volume 209 of *Proceedings of Machine Learning Research*, pages 578–597. PMLR.

Romain Lelong, Lina F Soualmia, Julien Grosjean, Mehdi Taalba, Stéfan J Darmoni, et al. 2019. Building a semantic health data warehouse in the context of clinical trials : development and usability study. *JMIR Medical Informatics*, 7(4) :e13917.

Paea LePendu, Srinivasan V Iyer, Anna Bauer-Mehren, Rave Harpaz, Jonathan M Mortensen, Tanya Podchiyska, Todd A Ferris, and Nigam H Shah. 2013. Pharmacovigilance using clinical notes. *Clinical pharmacology & therapeutics*, 93(6) :547–555.

Paea LePendu, Srinivasan V Iyer, Cédrick Fairon, and Nigam H Shah. 2012a. Annotation analysis for testing drug safety signals using unstructured clinical notes. In *Journal of biomedical semantics*, volume 3, pages 1–12. Springer.

Paea LePendu, Yi Liu, Srinivasan Iyer, Madeleine R Udell, and Nigam H Shah. 2012b. Analyzing patterns of drug use in clinical notes for patient safety. *AMIA Summits on Translational Science Proceedings*, 2012 :63.

Ivan Lerner, Jordan Jouffroy, Anita Burgun, and Antoine Neuraz. 2020a. Learning the grammar of prescription : recurrent neural network grammars for medication information extraction in clinical texts. *arXiv preprint ArXiv :2004.11622 [Cs].*

Ivan Lerner, Nicolas Paris, and Xavier Tannier. 2020b. Terminologies augmented recurrent neural network model for clinical named entity recognition. *Journal of biomedical informatics*, 102 :103356.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus : a resource for chemical disease relation extraction. *Database*, 2016.

Junyi Li, Xuejie Zhang, Xiaobing Zhou, et al. 2021a. Albert-based self-ensemble model with semisupervised learning and data augmentation for clinical semantic textual similarity calculation : algorithm validation study. *JMIR Medical Informatics*, 9(1) :e23086.

Li Li, Herbert S Chase, Chintan O Patel, Carol Friedman, and Chunhua Weng. 2008. Comparing icd9-encoded diagnoses and nlp-processed discharge summaries for clinical trials pre-screening : a case study. In *AMIA Annual Symposium Proceedings*, volume 2008, page 404. American Medical Informatics Association.

Minghao Li, Kyeryoung Lee, Zongzhi Liu, Meng Ma, Qi Pan, Rong Chen, Eric Schadt, and Xiaoyan Wang. 2021b. Applying bayesian hyperparameter optimization towards accurate and efficient topic modeling in clinical notes. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 493–494. IEEE.

Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan Pino, and Hassan Sajjad. 2019. Findings of the first shared task on machine translation robustness. *arXiv preprint arXiv :1906.11943*.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2023. A

comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association*, 30(2) :340–347.

Yikuan Li, Liang Yao, Chengsheng Mao, Anand Srivastava, Xiaoqian Jiang, and Yuan Luo. 2018. Early prediction of acute kidney injury in critical care setting using clinical notes. In *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 683–686. IEEE.

Ying Li, Hojjat Salmasian, Rave Harpaz, Herbert Chase, and Carol Friedman. 2011. Determining the reasons for medication prescriptions in the ehr using knowledge and natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2011, page 768. American Medical Informatics Association.

Chen Lin, Steven Bethard, Dmitriy Dligach, Farig Sadeque, Guergana Savova, and Timothy A Miller. 2020a. Does bert need domain adaptation for clinical negation detection ? *Journal of the American Medical Informatics Association*, 27(4) :584–591.

Wei-Chun Lin, Jimmy S Chen, Michael F Chiang, and Michelle R Hribar. 2020b. Applications of artificial intelligence to electronic health record data in ophthalmology. *Translational vision science & technology*, 9(2) :13–13.

Albee Y Ling, Allison W Kurian, Jennifer L Caswell-Jin, George W Sledge Jr, Nigam H Shah, and Suzanne R Tamang. 2019. Using natural language processing to construct a metastatic breast cancer cohort from linked cancer registry and electronic medical records data. *JAMIA open*, 2(4) :528–537.

Hongfang Liu, Stephen T Wu, Dingcheng Li, Siddhartha Jonnalagadda, Sunghwan Sohn, Kavishwar Wagholikar, Peter J Haug, Stanley M Huff, and Christopher G Chute. 2012a. Towards a semantic lexicon for clinical natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2012, page 568. American Medical Informatics Association.

Yi Liu, Paea LePendu, Srinivasan Iyer, and Nigam H Shah. 2012b. Using temporal patterns in medical records to discern adverse drug events from indications. *AMIA Summits on Translational Science proceedings*, 2012 :47.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

Tommaso Lo Barco, Mathieu Kuchenbuch, Nicolas Garcelon, Antoine Neuraz, and Rima Nabbout. 2021. Improving early diagnosis of rare diseases using natural language processing in unstructured medical records : an illustration from dravet syndrome. *Orphanet Journal of Rare Diseases*, 16 :1–12.

Sophia Loda, Jonathan Krebs, Sophia Danhof, Martin Schreder, Antonio G Solimando, Susanne Strifler, Leo Rasche, Martin Kortüm, Alexander Kerscher, Stefan Knop, et al. 2019. Exploration of artificial intelligence use with aries in multiple myeloma research. *Journal of Clinical Medicine*, 8(7) :999.

Henry J Lowe, Yang Huang, and Donald P Regula. 2009. Using a statistical natural language parser augmented with the umls specialist lexicon to assign snomed ct codes to anatomic sites and pathologic diagnoses in full text pathology reports. In *AMIA Annual Symposium Proceedings*, volume 2009, page 386. American Medical Informatics Association.

Qiuhao Lu, Dejing Dou, and Thien Nguyen. 2022. ClinicalT5: A generative language model for clinical text. In *Findings of the Association for Computational Linguistics : EMNLP 2022*, pages 5436–5443, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2022. Estimating the carbon footprint of bloom, a 176b parameter language model.

Stephen L Luther, Susan S Thomason, Sunil Sabharwal, Dezon K Finch, James McCart, Peter Toyinbo, Lina Bouayad, Michael E Matheny, Glenn T Gobbel, and Gail Powell-Cope. 2017. Leveraging electronic health care record information to measure pressure ulcer risk in veterans with spinal cord injury : a longitudinal study protocol. *JMIR research protocols*, 6(1) :e5948.

Kevin Lybarger, Meliha Yetisgen, and Özlem Uzuner. 2023. The 2022 n2c2/UW shared task on extracting social determinants of health. *Journal of the American Medical Informatics Association*.

Julia Madec, Guillaume Bouzillé, Christine Riou, Pascal Van Hille, Christian Merour, Marie-Lisen Artigny, Denis Delamarre, Veronique Raimbert, Pierre Lemordant, and Marc Cuggia. 2019. ehop clinical data warehouse: From a prototype to the creation

of an inter-regional clinical data centers network. *Studies in health technology and informatics*, 264 :1536—1537.

Sanne Magnan. 2017. Social determinants of health 101 for health care : five plus five. *NAM perspectives*.

Christopher J Magnani, Nicolas Bievre, Laurence C Baker, James D Brooks, Douglas W Blayney, and Tina Hernandez-Boussard. 2021. Real-world evidence to estimate prostate cancer costs for first-line treatment or active surveillance. *European urology open science*, 23 :20–29.

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. 2020. The e3c project : Collection and annotation of a multilingual corpus of clinical cases. *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020*.

Diwakar Mahajan, Ananya Poddar, Jennifer J Liang, Yen-Ting Lin, John M Prager, Parthasarathy Suryanarayanan, Preethi Raghavan, Ching-Huei Tsou, et al. 2020. Identification of semantically similar sentences in clinical notes : Iterative intermediate training using multi-task learning. *JMIR medical informatics*, 8(11) :e22508.

Scott A Malec, Peng Wei, Elmer V Bernstam, Richard D Boyce, and Trevor Cohen. 2021. Using computable knowledge mined from the literature to elucidate confounders for ehr-based pharmacovigilance. *Journal of biomedical informatics*, 117 :103719.

François Maniez. 2011. L'apport des corpus spécialisés en terminographie multilingue : le cas des groupes nominaux de type nom-adjectif dans la langue médicale. *Meta*, 56(2) :391–406.

Michael Marmot. 2005. Social determinants of health inequalities. *The lancet*, 365(9464) :1099–1104.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Genevieve B Melton, Simon Parsons, Frances P Morrison, Adam S Rothschild, Marianthi Markatou, and George Hripcsak. 2006. Inter-patient distance metrics using snomed ct defining relationships. *Journal of biomedical informatics*, 39(6) :697–705.

S M Meystre, C Lovis, T Bürkle, G Tognola, A Budrionis, and C U Lehmann. 2017. Clinical data reuse or secondary use : Current status and potential future progress. *Yearb. Med. Inform.*, 26(1) :38–52.

Stéphane M Meystre, Paul M Heider, Youngjun Kim, Daniel B Aruch, and Carolyn D Britten. 2019. Automatic trial eligibility surveillance based on unstructured clinical data. *International journal of medical informatics*, 129 :13–19.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Diamos, Erich Elsen, David García, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2017. Mixed precision training. *CoRR*, abs/1710.03740.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Taejin L Min, Liyan Xu, Jinho D Choi, Ranliang Hu, Jason W Allen, Christopher Reeves, Derek Hsu, Richard Duszak Jr, Jeffrey Switchenko, and Gelareh Sadigh. 2022. Covid-19 pandemic-associated changes in the acuity of brain mri findings : A secondary analysis of reports using natural language processing. *Current Problems in Diagnostic Radiology*, 51(4) :529–533.

Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, and Martin Krallinger. 2020. Overview of automatic clinical coding : Annotations, guidelines, and solutions for non-english clinical cases at codiesp track of clef ehealth 2020. *CLEF (Working Notes)*, 2020.

Mehdi Mirzapour, Amine Abdaoui, Andon Tchechmedjiev, William Digan, Sandra Bringay, and Clement Jonquet. 2021. French fastcontext : A publicly accessible system for detecting negation, temporality and experiencer in french clinical notes. *Journal of Biomedical Informatics*, 117 :103733.

David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, and PRISMA Group*. 2009. Preferred reporting items for systematic reviews and meta-analyses : the prisma statement. *Annals of internal medicine*, 151(4) :264–269.

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

George B Moody and Roger G Mark. 1996. A database to support development and evaluation of intelligent intensive care monitoring. In *Computers in Cardiology 1996*, pages 657–660. IEEE.

Sungrim Moon, Donna Ihrke, Yuqun Zeng, and Hongfang Liu. 2017. Distinction between medical and non-medical usages of short forms in clinical narratives. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1302. American Medical Informatics Association.

Sungrim Moon, Andrew Wen, Christopher G Scott, Peter A Noseworthy, Jeffrey G Geske, Jane L Shellum, Rajeev Chaudhry, Steve R Ommen, Rick A Nishimura, Hongfang Liu, et al. 2018. An automated system for analysis of implantable cardioverter defibrillator reports in hypertrophic cardiomyopathy patients. *Circulation*, 138(Suppl_1) :A16215–A16215.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning.

Faith W Mutinda, Sumaila Nigo, Daisaku Shibata, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2020. Detecting redundancy in electronic medical records using clinical bert. In *Proceedings of the 26th Annual Conference of the Association for Natural Language Processing (NLP2020), Online*, pages 16–19.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https ://github.com/chakki-works/seqeval.

Antoine Neuraz, Ivan Lerner, William Digan, Nicolas Paris, Rosy Tsopra, Alice Rogier, David Baudoin, Kevin Bretonnel Cohen, Anita Burgun, Nicolas Garcelon, et al. 2020.

Natural language processing for rapid response to emergent diseases : case study of calcium channel blockers and hypertension in the covid-19 pandemic. *Journal of medical Internet research*, 22(8) :e20773.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english : opportunities and challenges. *Journal of biomedical semantics*, 9(1) :1–13.

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The quaero french medical corpus : A ressource for medical entity recognition and normalization. *Proc of BioTextMining Work*, pages 24–30.

Mariana Neves and Jurica Ševa. 2021. An extensive review of tools for manual annotation of documents. *Briefings in bioinformatics*, 22(1) :146–163.

Sagar U Nigwekar, Craig A Solid, Elizabeth Ankers, Rajeev Malhotra, William Eggert, Alexander Turchin, Ravi I Thadhani, and Charles A Herzog. 2014. Quantifying a rare disease in administrative data : the example of calciphylaxis. *Journal of general internal medicine*, 29 :724–731.

OpenAI. 2023. Gpt-4 technical report.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

John D Osborne, Matthew Wyatt, Andrew O Westfall, James Willig, Steven Bethard, and Geoff Gordon. 2016. Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning. *Journal of the American Medical Informatics Association*, 23(6) :1077–1084.

Casey Lynnette Overby, Jyotishman Pathak, Omri Gottesman, Krystl Haerian, Adler Perotte, Sean Murphy, Kevin Bruce, Stephanie Johnson, Jayant Talwalkar, Yufeng Shen, et al. 2013. A collaborative approach to developing an electronic health record phenotyping algorithm for drug-induced liver injury. *Journal of the American Medical Informatics Association*, 20(e2) :e243–e252.

Evangelos Pafilis, Sune P Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Aikaterini Vasileiadou, Christos Arvanitidis, and Lars Juhl Jensen. 2013. The species and organisms resources for fast and accurate identification of taxonomic names in text. *PloS one*, 8(6) :e65390.

Serguei Pakhomov, Bridget McInnes, Terrence Adam, Ying Liu, Ted Pedersen, and Genevieve B Melton. 2010. Semantic similarity and relatedness between clinical terms : an experimental study. In *AMIA annual symposium proceedings*, volume 2010, page 572. American Medical Informatics Association.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa : A large corpus for question answering on electronic medical records. *arXiv preprint arXiv :1809.00732*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Tejal A Patel, Mamta Puppala, Richard O Ogunti, Joe E Ensor, Tiancheng He, Jitesh B Shewale, Donna P Ankerst, Virginia G Kaklamani, Angel A Rodriguez, Stephen TC Wong, et al. 2017. Correlating mammographic and pathologic findings in clinical decision support using natural language processing and data mining methods. *Cancer*, 123(1) :114–121.

Braja G Patra, Mohit M Sharma, Veer Vekaria, Prakash Adekkanattu, Olga V Patterson, Benjamin Glicksberg, Lauren A Lepow, Euijung Ryu, Joanna M Biernacka, Al'ona Furmanchuk, Thomas J George, William Hogan, Yonghui Wu, Xi Yang, Jiang Bian, Myrna Weissman, Priya Wickramaratne, J John Mann, Mark Olfson, Jr Campion, Thomas R, Mark Weiner, and Jyotishman Pathak. 2021. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *Journal of the American Medical Informatics Association*, 28(12) :2716–2727.

Jon D Patrick, Dung HM Nguyen, Yefeng Wang, and Min Li. 2011. A knowledge discovery

and reuse pipeline for information extraction in clinical notes. *Journal of the american medical informatics association*, 18(5) :574–579.

Ted Pedersen, Serguei VS Pakhomov, Siddharth Patwardhan, and Christopher G Chute. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics*, 40(3) :288–299.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019a. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019b. Transfer learning in biomedical natural language processing : an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv :1906.05474*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014a. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014b. Glove : Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Adler Perotte, Rajesh Ranganath, Jamie S Hirsch, David Blei, and Noémie Elhadad. 2015. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. *Journal of the American Medical Informatics Association*, 22(4) :872–880.

Adler Perotte, Frank Wood, Noemie Elhadad, and Nicholas Bartlett. 2011. Hierarchically supervised latent dirichlet allocation. *Advances in neural information processing systems*, 24.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long*

*Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Anne-Dominique Pham, Aurélie Névéol, Thomas Lavergne, Daisuke Yasunaga, Olivier Clément, Guy Meyer, Rémy Morello, and Anita Burgun. 2014. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. *BMC bioinformatics*, 15(1) :1–10.

Long N Phan, James T Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive : a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv :2106.03598*.

Rimma Pivovarov and Noémie Elhadad. 2012. A hybrid knowledge-based and data-driven approach to identifying semantically similar concepts. *Journal of biomedical informatics*, 45(3) :471–481.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. 2018. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1) :180178.

Thibaut Pressat-Laffouilhère, Pierre Balayé, Badisse Dahamna, Romain Lelong, Kévin Billey, Stéfan J Darmoni, and Julien Grosjean. 2022. Evaluation of doc'eds : a french semantic search tool to query health documents from a clinical data warehouse. *BMC medical informatics and decision making*, 22(1) :34.

Sampo Pyysalo, Tomoko Ohta, and Jun'ichi Tsujii. 2011. Overview of the entity relations (rel) supporting task of bionlp shared task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, BioNLP Shared Task '11, page 83–88, USA. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8) :9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Ali S Raja, Sarvenaz Pourjabbar, Ivan K Ip, Christopher W Baugh, Aaron D Sodickson, Michael O'Leary, and Ramin Khorasani. 2019. Impact of a health information technology–enabled appropriate use criterion on utilization of emergency department ct for renal colic. *American Journal of Roentgenology*, 212(1) :142–145.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Joseph S Redman, Yamini Natarajan, Jason K Hou, Jingqi Wang, Muzammil Hanif, Hua Feng, Jennifer R Kramer, Roxanne Desiderio, Hua Xu, Hashem B El-Serag, et al. 2017. Accurate identification of fatty liver disease in data warehouse utilizing natural language processing. *Digestive diseases and sciences*, 62 :2713–2718.

Ruth M Reeves, Lee Christensen, Jeremiah R Brown, Michael Conway, Maxwell Levis, Glenn T Gobbel, Rashmee U Shah, Christine Goodrich, Iben Ricket, Freneka Minter, Andrew Bohm, Bruce E Bray, Michael E Matheny, and Wendy Chapman. 2021. Adaptation of an NLP system to a new healthcare environment to identify social determinants of health. *J. Biomed. Inform.*, 120(103851) :103851.

Robert A Riestenberg, Andrew Furman, Avery Cowen, Anna Pawlowksi, Daniel Schneider, Alana A Lewis, Sean Kelly, Babafemi Taiwo, Chad Achenbach, Frank Palella, et al. 2019. Differences in statin utilization and lipid lowering by race, ethnicity, and hiv status in a real-world cohort of persons with human immunodeficiency virus and uninfected persons. *American heart journal*, 209 :79–87.

Kirk Roberts, Yuqi Si, Anshul Gandhi, and Elmer Bernstam. 2018. A framenet for cancer information in clinical narratives : Schema and annotation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Christian M Rochefort, David L Buckeridge, and Michal Abrahamowicz. 2015. Improving patient safety by optimizing the use of nursing human resources. *Implementation Science*, 10 :1–11.

Christian M Rochefort, David L Buckeridge, Andréanne Tanguay, Alain Biron, Frédérick D'Aragon, Shengrui Wang, Benoit Gallix, Louis Valiquette, Li-Anne Audet, Todd C

Lee, et al. 2017. Accuracy and generalizability of using automated methods for identifying adverse events from electronic health record data : a validation study protocol. *BMC Health Services Research*, 17(1) :1–9.

Alice Rogier, Adrien Coulet, and Bastien Rance. 2021. Using an ontological representation of chemotherapy toxicities for guiding information extraction and integration from ehrs. In *Medinfo 2021-18th World Congress on Medical and Health Informatics*.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain.

Justine H Ryu and Andrew J Zimolzak. 2020. Natural language processing of serum protein electrophoresis reports in the veterans affairs health care system. *JCO Clinical Cancer Informatics*, 4 :749–756.

Charles Safran, Meryl Bloomrosen, W. Edward Hammond, Steven Labkoff, Suzanne Markel-Fox, Paul C. Tang, and Don E. Detmer. 2007. Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. *Journal of the American Medical Informatics Association*, 14(1) :1–9.

Tabinda Sarwar, Sattar Seifollahi, Jeffrey Chan, Xiuzhen Zhang, Vural Aksakalli, Irene Hudson, Karin Verspoor, and Lawrence Cavedon. 2022. The secondary use of electronic health records for data mining : Data characteristics and challenges. *ACM Computing Surveys (CSUR)*, 55(2) :1–40.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom : A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv :2211.05100*.

Elyne Scheurwegs, Kim Luyckx, Léon Luyten, Walter Daelemans, and Tim Van den Bulcke. 2016. Data integration of structured and unstructured sources for assigning clinical codes to patient stays. *Journal of the American Medical Informatics Association*, 23(e1) :e11–e19.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro

Barra. 2020. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

Claudia Schulz, Josh Levy-Kramer, Camille Van Assel, Miklos Kepes, and Nils Hammerla. 2020. Biomedical concept relatedness – a large EHR-based benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6565–6575, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

William A Scott. 1955. Reliability of content analysis : The case of nominal scale coding. *Public opinion quarterly*, pages 321–325.

Amikar Sehdev, Ross Hayden, Mathew Joseph Kuhar, Liang Cheng, Simon John Warren, Lawrence Aaron Mark, William Arthur Wooden, Douglas Jay Schwartzentruber, and Theodore F Logan. 2018. Prognostic role of braf mutation in malignant cutaneous melanoma.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shahid Munir Shah and Rizwan Ahmed Khan. 2020. Secondary use of electronic health record : Opportunities and challenges. *IEEE access*, 8 :136947–136965.

Yijun Shao, Qing T Zeng, Kathryn K Chen, Andrew Shutes-David, Stephen M Thielke, and Debby W Tsuang. 2019. Detection of probable dementia cases in undiagnosed patients using structured and unstructured electronic health records. *BMC medical informatics and decision making*, 19(1) :1–11.

Brihat Sharma, Dmitriy Dligach, Kristin Swope, Elizabeth Salisbury-Afshar, Niranjan S Karnik, Cara Joyce, and Majid Afshar. 2020. Publicly available machine learning models for identifying opioid misuse from the clinical notes of hospitalized patients. *BMC medical informatics and decision making*, 20 :1–11.

Camara Sharperson, Tarek N Hanna, Keith D Herr, Matthew E Zygmont, Roger L Gerard, and Jamlik-Omari Johnson. 2021. The effect of covid-19 on emergency department imaging : what can we learn ? *Emergency Radiology*, 28 :339–347.

Yoshie Shimai, Toshihiro Takeda, Katsuki Okada, Shirou Manabe, Kei Teramoto, Naoki Mihara, and Yasushi Matsumura. 2018. Screening of anticancer drugs to detect drug-induced interstitial pneumonia using the accumulated data in the electronic medical record. *Pharmacology Research & Perspectives*, 6(4) :e00421.

Kimberly Shoenbill, Yiqiang Song, Mark Craven, Heather Johnson, Maureen Smith, and Eneida A Mendonca. 2020. Identifying patterns and predictors of lifestyle modification in electronic health record documentation using statistical and machine learning methods. *Preventive medicine*, 136 :106061.

Karandeep Singh, Rebecca A Betensky, Adam Wright, Gary C Curhan, David W Bates, and Sushrut S Waikar. 2016. A concept–wide association study of clinical notes to discover new predictors of kidney failure. *Clinical Journal of the American Society of Nephrology*, 11(12) :2150–2158.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023a. Large language models encode clinical knowledge. *Nature*, pages 1–9.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023b. Towards expert-level medical question answering with large language models.

Larry Smith, Lorraine K Tanabe, Cheng-Ju Kuo, I Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, Manabu Torii, et al. 2008. Overview of biocreative ii gene mention recognition. *Genome biology*, 9(2) :1–19.

Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. 2017. Biosses : a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14) :i49–i58.

Illés Solt, Domonkos Tikk, Viktor Gál, and Zsolt T Kardkovács. 2009. Semantic classification of diseases in discharge summaries using a context-aware rule-based classifier. *Journal of the American Medical Informatics Association*, 16(4) :580–584.

Hyunju Song, Yang Gu, Gondy Leroy, Fariba M Donovan, and John N Galgiani. 2021. Integrating automated biomedical lexicon creation for valley fever diagnosis. In *2021 IEEE/ACM Conference on Connected Health : Applications, Systems and Engineering Technologies (CHASE)*, pages 111–112. IEEE.

Rachel Stemerman, Jaime Arguello, Jane Brice, Ashok Krishnamurthy, Mary Houston, and Rebecca Kitzmiller. 2021. Identification of social determinants of health using multi-label classification of electronic health record clinical notes. *JAMIA open*, 4(3) :ooaa069.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat : a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Reejis Stephen, Aziz Boxwala, and Paul Gertman. 2003. Feasibility of using a large clinical data warehouse to automate the selection of diagnostic cohorts. In *AMIA Annual Symposium Proceedings*, volume 2003, page 1019. American Medical Informatics Association.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives : Overview of 2014 i2b2/uthealth shared task track 1. *Journal of biomedical informatics*, 58 :S11–S19.

Mengxin Sun, Jason Baron, Anand Dighe, Peter Szolovits, Richard G Wunderink, Tamara Isakova, and Yuan Luo. 2019. Early prediction of acute kidney injury in critical care setting using clinical notes and structured multivariate physiological measurements. *MedInfo*, 264 :368–372.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, 20(5) :806–813.

Suzanne Tamang, Manali I Patel, Douglas W Blayney, Julie Kuznetsov, Samuel G Finlayson, Yohan Vetteth, and Nigam Shah. 2015. Detecting unplanned care from clinician notes in electronic health records. *Journal of Oncology Practice*, 11(3) :e313–e319.

Yi Tay, Dara Bahri, Liu Yang, Donald Metzler, and Da-Cheng Juan. 2020. Sparse sinkhorn attention. In *International Conference on Machine Learning*, pages 9438–9447. PMLR.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, and Donald Metzler. 2022. Unifying language learning paradigms. *arXiv preprint arXiv :2205.05131*.

Patrick J Thoral, Jan M Peppink, Ronald H Driessen, Eric JG Sijbrands, Erwin JO Kompanje, Lewis Kaplan, Heatherlee Bailey, Jozef Kesecioglu, Maurizio Cecconi, Matthew Churpek, et al. 2021. Sharing icu patient data responsibly under the society of critical care medicine/european society of intensive care medicine joint data science collaboration : the amsterdam university medical centers database (amsterdamumcdb) example. *Critical care medicine*, 49(6) :e563.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel and free: `http://logos.uio.no/opus`. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

M Tien, R Kashyap, GA Wilson, V Hernandez-Torres, AK Jacob, DR Schroeder, and Carlos B Mantilla. 2015. Retrospective derivation and validation of an automated electronic search algorithm to identify post operative cardiovascular and thromboembolic complications. *Applied clinical informatics*, 6(03) :565–576.

Daniel To, Brihat Sharma, Niranjan Karnik, Cara Joyce, Dmitriy Dligach, and Majid Afshar. 2020. Validation of an alcohol misuse classifier in hospitalized patients. *Alcohol*, 84 :49–55.

Rian Touchent, Laurent Romary, and Eric de la Clergerie. 2023. Camembert-bio: a tasty french language model better for your health.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2 : Open foundation and fine-tuned chat models. *arXiv preprint arXiv :2307.09288*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1) :1–28.

Hazal Turkmen, Oguz Dikenelli, Cenk Eraslan, and Mehmet Cem Callı. 2022. Bioberturk : Exploring turkish biomedical language model development strategies in low resource setting. *Preprint from Research Square*.

Sudhi G Upadhyaya, Dennis H Murphree Jr, Che G Ngufor, Alison M Knight, Daniel J Cronk, Robert R Cima, Timothy B Curry, Jyotishman Pathak, Rickey E Carter, and Daryl J Kor. 2017. Automated diabetes case identification using electronic health record data at a tertiary care facility. *Mayo Clinic Proceedings : Innovations, Quality & Outcomes*, 1(1) :100–110.

Özlem Uzuner. 2009. Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association*, 16(4) :561–570.

Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R South. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*, 19(5) :786–791.

Ozlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2008a. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15(1) :14–24.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5) :514–518.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5) :552–556.

Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2008b. Identifying Patient Smoking Status from Medical Discharge Records. *Journal of the American Medical Informatics Association*, 15(1) :14–24.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the State-of-the-Art in Automatic De-identification. *Journal of the American Medical Informatics Association*, 14(5) :550–563.

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Downstream task performance of BERT models pre-trained using automatically de-identified clinical data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4245–4252, Marseille, France. European Language Resources Association.

Erik M Van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A Kors, and Laura I Furlong. 2012. The eu-adr corpus : annotated drugs, diseases, targets, and their relationships. *Journal of biomedical informatics*, 45(5) :879–884.

Tielman T Van Vleck, Daniel M Stein, Peter D Stetson, and Stephen B Johnson. 2007. Assessing data relevance for automated generation of a clinical summary. In *AMIA annual symposium proceedings*, volume 2007, page 761. American Medical Informatics Association.

Tielman T Van Vleck, Adam Wilcox, Peter D Stetson, Stephen B Johnson, and Noémie Elhadad. 2008. Content and structure of clinical problem lists : a corpus analysis. In *AMIA Annual Symposium Proceedings*, volume 2008, page 753. American Medical Informatics Association.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Sumithra Velupillai, Hanna Suominen, Maria Liakata, Angus Roberts, Anoop D Shah, Katherine Morley, David Osborn, Joseph Hayes, Robert Stewart, Johnny Downs, Wendy Chapman, and Rina Dutta. 2018. Using clinical natural language processing for health outcomes research : Overview and actionable suggestions for future advances. *J. Biomed. Inform.*, 88 :11–19.

A. Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2) :260–269.

Jessica A Walsh, Yijun Shao, Jianwei Leng, Tao He, Chia-Chen Teng, Doug Redd, Qing Treitler Zeng, Zachary Burningham, Daniel O Clegg, and Brian C Sauer. 2017. Identifying axial spondyloarthritis in electronic medical records of us veterans. *Arthritis care & research*, 69(9) :1414–1420.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue : A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018a. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018b. Glue : A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv :1804.07461*.

Guan Wang, Kenneth Jung, Rainer Winnenburg, and Nigam H Shah. 2015a. A method for systematic discovery of adverse drug events from clinical notes. *Journal of the American Medical Informatics Association*, 22(6) :1196–1204.

Lin Wang, Zhong Xue, Chika F Ezeana, Mamta Puppala, Shenyi Chen, Rebecca L Danforth, Xiaohui Yu, Tiancheng He, Mark L Vassallo, and Stephen TC Wong. 2019b. Preventing inpatient falls with injuries using integrative machine learning prediction : A cohort study. *NPJ digital medicine*, 2(1) :127.

Liwei Wang, Jason Wampfler, Angela Dispenzieri, Hua Xu, Ping Yang, and Hongfang Liu. 2019c. Achievability to extract specific date information for cancer research. In *AMIA Annual Symposium Proceedings*, volume 2019, page 893. American Medical Informatics Association.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer : Self-attention with linear complexity. *arXiv preprint arXiv :2006.04768*.

Xiaoyan Wang, Herbert Chase, Marianthi Markatou, George Hripcsak, and Carol Friedman. 2010. Selecting information in electronic health records for knowledge acquisition. *Journal of biomedical informatics*, 43(4) :595–601.

Xiaoyan Wang, Amy Chused, Noémie Elhadad, Carol Friedman, and Marianthi Markatou. 2008. Automated knowledge acquisition from clinical narrative reports. In *AMIA Annual Symposium Proceedings*, volume 2008, page 783. American Medical Informatics Association.

Xiaoyan Wang, George Hripcsak, Marianthi Markatou, and Carol Friedman. 2009. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records : a feasibility study. *Journal of the American Medical Informatics Association*, 16(3) :328–337.

Yan Wang, Elizabeth S Chen, Serguei Pakhomov, Elliot Arsoniadis, Elizabeth W Carter, Elizabeth Lindemann, Indra Neil Sarkar, and Genevieve B Melton. 2015b. Automated extraction of substance use information from clinical texts. *AMIA Annu. Symp. Proc.*, 2015 :2121–2130.

Yanshan Wang, Saeed Mehrabi, Sunghwan Sohn, Elizabeth J Atkinson, Shreyasee Amin, and Hongfang Liu. 2019d. Natural language processing of radiology reports for identification of skeletal site-specific fractures. *BMC Medical Informatics and Decision Making*, 19 :23–29.

Matthieu Wargny. 2022. *Diabète et insuffisance cardiaque : approche épidémiologique par l'analyse croisée de différentes sources de données de santé*. Theses, Nantes Université.

Hannah L Weeks, Cole Beck, Elizabeth McNeer, Michael L Williams, Cosmin A Bejan, Joshua C Denny, and Leena Choi. 2020. medextractr : A targeted, customizable approach to medication extraction from electronic health records. *Journal of the American Medical Informatics Association*, 27(3) :407–418.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L Gwet. 2013. A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients : a study conducted with personality disorder samples. *BMC medical research methodology*, 13 :1–7.

A Wright, A McCoy, S Henkin, M Flaherty, and D Sittig. 2013. Validation of an association rule mining-based method to infer associations between medications and problems. *Applied Clinical Informatics*, 4(01) :100–109.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Further finetuning llama on medical papers.

David Wei Wu, Jonathan A Bernstein, and Gill Bejerano. 2022. Discovering monogenic patients with a confirmed molecular diagnosis in millions of clinical notes with monominer. *Genetics in Medicine*, 24(10) :2091–2102.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

Alexandre Yahi and Nicholas P Tatonetti. 2015. A knowledge-based, automated method for phenotyping in the ehr using only clinical pathology reports. *AMIA Summits on Translational Science Proceedings*, 2015 :64.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. 2022. A large language model for electronic health records. *npj Digital Medicine*, 5(1) :194.

Xi Yang, Tianchen Lyu, Qian Li, Chih-Yin Lee, Jiang Bian, William R Hogan, and Yonghui Wu. 2019. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC medical informatics and decision making*, 19(5) :1–9.

Xi Yang, Hanyuan Yang, Tianchen Lyu, Shuang Yang, Yi Guo, Jiang Bian, Hua Xu, and Yonghui Wu. 2020a. A natural language processing tool to extract quantitative smoking status from clinical narratives. In *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–2. IEEE.

Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020b. Finbert: A pretrained language model for financial communications.

Meliha Yetisgen, Elena Pellicer, David R Crosslin, and Lucy Vanderwende. 2016. Automatic identification of lifestyle and environmental factors from social history in clinical text. In *CRI*.

Meliha Yetisgen and Lucy Vanderwende. 2017. Automatic identification of substance abuse from social history in clinical text. In *Artificial Intelligence in Medicine*, pages 171–181, Cham. Springer International Publishing.

Zehao Yu, Xi Yang, Chong Dang, Songzi Wu, Prakash Adekkanattu, Jyotishman Pathak, Thomas J George, William R Hogan, Yi Guo, Jiang Bian, and Yonghui Wu. 2021a. A study of social and behavioral determinants of health in lung cancer patients using transformers-based natural language processing models. *AMIA Annu. Symp. Proc.*, 2021 :1225–1233.

Zehao Yu, Xi Yang, Chong Dang, Songzi Wu, Prakash Adekkanattu, Jyotishman Pathak, Thomas J George, William R Hogan, Yi Guo, Jiang Bian, et al. 2021b. A study of social and behavioral determinants of health in lung cancer patients using transformers-based natural language processing models. In *AMIA Annual Symposium Proceedings*, volume 2021, page 1225. American Medical Informatics Association.

Zehao Yu, Xi Yang, Yi Guo, Jiang Bian, and Yonghui Wu. 2022. Assessing the documentation of social determinants of health for lung cancer patients in clinical narratives. *Front. Public Health*, 10 :778463.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. Biobart : Pretraining and evaluation of a biomedical generative language model. *arXiv preprint arXiv :2204.03905*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird : Transformers for longer sequences. *Advances in neural information processing systems*, 33 :17283–17297.

Qing T Zeng, Doug Redd, Thomas Rindflesch, and Jonathan Nebeker. 2012. Synonym, topic model and predicate-based query expansion for retrieving clinical documents. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1050. American Medical Informatics Association.

Xianghao Zhan, Marie Humbert-Droz, Pritam Mukherjee, and Olivier Gevaert. 2021. Structuring clinical text with ai : Old versus new natural language processing techniques evaluated on eight common cardiovascular diseases. *Patterns*, 2(7) :100289.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2022. CBLUE: A Chinese biomedical language

understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 7888–7915, Dublin, Ireland. Association for Computational Linguistics.

Yian Zhang, Alex Warstadt, Haau-Sing Li, and Samuel R. Bowman. 2020. When do you need billions of words of pretraining data?

Yiqing Zhao, Anastasios Dimou, Feichen Shen, Nansu Zong, Jaime I Davila, Hongfang Liu, and Chen Wang. 2022. Po2rdf : representation of real-world data for precision oncology using resource description framework. *BMC medical genomics*, 15(1) :1–12.

Yiqing Zhao, Saravut J Weroha, Ellen L Goode, Hongfang Liu, and Chen Wang. 2021. Generating real-world evidence from unstructured clinical notes to examine clinical utility of genetic tests : use case in brcaness. *BMC Medical Informatics and Decision Making*, 21 :1–13.

Qiu-Yue Zhong, Elizabeth W Karlson, Bizu Gelaye, Sean Finan, Paul Avillach, Jordan W Smoller, Tianxi Cai, and Michelle A Williams. 2018. Screening pregnant women for suicidal behavior in electronic medical records : diagnostic codes vs. clinical notes processed by natural language processing. *BMC medical informatics and decision making*, 18(1) :1–11.

Li Zhou, Genevieve B Melton, Simon Parsons, and George Hripcsak. 2006. A temporal constraint structure for extracting temporal information from clinical narrative. *Journal of biomedical informatics*, 39(4) :424–439.

Xin Zhou, Yanshan Wang, Sunghwan Sohn, Terry M Therneau, Hongfang Liu, and David S Knopman. 2019. Automatic extraction and assessment of lifestyle exposures for alzheimer's disease using natural language processing. *International journal of medical informatics*, 130 :103943.

Dongqing Zhu, Stephen Wu, Ben Carterette, and Hongfang Liu. 2014. Using large clinical corpora for query expansion in text-based cohort identification. *Journal of biomedical informatics*, 49 :275–281.

Hongyin Zhu, Hao Peng, Zhiheng Lyu, Lei Hou, Juanzi Li, and Jinghui Xiao. 2021. Travelbert: Pre-training language model incorporating domain-specific heterogeneous knowledge into a unified representation.

Vivienne J Zhu, Leslie A Lenert, Brian E Bunnell, Jihad S Obeid, Melanie Jefferson, and Chanita Hughes Halbert. 2019. Automatically identifying social isolation from clinical narratives for patients with prostate cancer. *BMC medical informatics and decision making*, 19(1) :1–9.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies : Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Nansu Zong, Victoria Ngo, Daniel J Stone, Andrew Wen, Yiqing Zhao, Yue Yu, Sijia Liu, Ming Huang, Chen Wang, Guoqian Jiang, et al. 2021. Leveraging genetic reports and electronic health records for the prediction of primary cancers : algorithm development and validation study. *JMIR Medical Informatics*, 9(5) :e23586.

Maximilian Zubke, Matthias Katzensteiner, and Oliver J Bott. 2020. Integration of unstructured data into a clinical data warehouse for kidney transplant screening-challenges & solutions. In *MIE*, pages 272–276.

Xu Zuo, Jianfu Li, Bo Zhao, Yujia Zhou, Xiao Dong, Jon Duke, Karthik Natarajan, George Hripcsak, Nigam Shah, Juan M Banda, et al. 2020. Normalizing clinical document titles to loinc document ontology : An initial study. In *AMIA Annual Symposium Proceedings*, volume 2020, page 1441. American Medical Informatics Association.

$<$

COLLEGE DOCTORAL PAYS DE LA LOIRE ED MATHEMATIQUES STIC

Nantes Université

**Titre :** TALMed : Traitement Automatique de la Langue Médicale

**Mot clés :** TAL clinique, modèle de langue pré-entrainés, entrepôts de données de santé

**Résumé :** La collecte massive de données de santé a permis l'émergence d'usages secondaires, notamment la recherche et l'évaluation de la qualité des soins. Pour une utilisation optimale, ces données doivent être harmonisées et stockées dans des entrepôts de données de santé (EDS), souvent sous forme textuelle. Le traitement automatique des langues (TAL) est alors nécessaire pour en extraire des informations à grande échelle. Les méthodes actuelles de TAL s'appuient principalement sur des modèles de langue basés sur l'architecture Transformer, qui nécessitent d'être adaptés au domaine médical pour tirer profit du potentiel de ces modèles. Dans cette thèse, nous explorons deux thématiques : l'adaptation de ces modèles au contexte médical français et leur application en recherche clinique. Premièrement, nous menons plusieurs études d'adaptation au domaine médical de différents modèles pré-entraînés existants. Ces études ont pour but d'évaluer l'impact de différents paramètres pour l'adaptation des modèles, comme la nature des données ou la stratégie de pré-entraînement. Enfin, l'utilisation de ces modèles est étudiée dans deux projets de recherche clinique. Le projet GAVROCHE examine la relation entre la variabilité glycémique et la mortalité chez les patients atteints d'insuffisance cardiaque aiguë. Le second projet vise à extraire des déterminants sociaux de santé à partir des comptes rendus cliniques. Ces cas montrent le potentiel du TAL pour extraire des informations cliniques cruciales.

**Title:** Medical Natural Language Processing

**Keywords:** clinical NLP, pretrained language models, clinical data warehouse

**Abstract:** The massive collection of health data has allowed the emergence of secondary uses, including research and the evaluation of the quality of care. For optimal use, these data need to be harmonized and stored in health data warehouses (HDWs), often in textual form. Natural Language Processing (NLP) is then required to extract information on a large scale. Current NLP methods mainly rely on language models based on the Transformer architecture, which need to be adapted to the medical field to benefit from the potential of these models. In this thesis, we explore two themes: the adaptation of these models to the French medical context and their application in clinical research. First, we conduct several studies on the adaptation of various pre-existing pre-trained models to the medical field. The aim of these studies is to evaluate the impact of different parameters for model adaptation, such as the nature of the data or the pre-training strategy. Finally, the use of these models is studied in two clinical research projects. The GAVROCHE project examines the relationship between glycemic variability and mortality in patients with acute heart failure. The second project aims to extract social health determinants from clinical reports. These cases demonstrate the potential of NLP to extract crucial clinical information.