**UNIVERSITE DE BLIDA 1**
**Faculté des sciences**
Département d'informatique

**MEMOIRE DE MASTER**
**En Informatique**

Option : Ingénierie du Logiciel

# Une Approche Sémantique Pour La Simplification Automatique Des Textes Médicaux

Réalisé par
Mlle. LAOUICHI Zineb
Mlle. GUERBA Khaoula

Encadré par
Mme. MEZZI Melyara

En présence des membres du jury

| | | |
|---|---|---|
| **Présidente :** | Dr. FAREH | Université De Blida 1 |
| **Examinatrice :** | Dr. CHERIGUENE | Université De Blida 1 |

Juillet 2023

Résumé

Ce mémoire présente une approche sémantique visant à simplifier les textes médicaux. Trois solutions distinctes sont mises en œuvre dans cette approche; Dans la première solution, le modèle pré-entraîné BioGPT est utilisé pour générer du texte médical. La deuxième solution repose sur l'algorithme TextRank, qui permet de résumer les textes médicaux. Enfin, la dernière solution exploite les ontologies de UMLS et la ressource lexico-semantique WordNet afin de simplifier le langage médical, complétant ainsi les deux premières solutions.

Ce travail de recherche apporte une contribution significative au domaine de la simplification des textes médicaux en proposant des solutions concrètes. L'approche sémantique adoptée offre des perspectives prometteuses pour faciliter la compréhension et l'utilisation des informations médicales. Une telle simplification des textes médicaux peut avoir un impact positif sur la qualité des soins de santé et favoriser une meilleure accessibilité aux informations médicales pour une large gamme de lecteurs. En résumé, ce mémoire souligne l'importance de cette approche et ses implications potentielles dans le domaine de la santé.

**Mots-clés:** Traitement Automatique du Langage Naturel, Génération Automatique de Texte, Résumé de Texte, Simplification des Textes, Analyse Sémantique.

# الملخص

هذه الأطروحة العلمية تقدم نهجًا معنويًا يهدف إلى تبسيط النصوص الطبية. يتم تنفيذ ثلاث حلول متميزة ضمن هذا النهج حيث يستخدم الحل الأول النموذج المدرب مسبقًا (BioGPT) لإنتاج النص الطبي. بينما يعتمد الحل الثاني على خوارزمية (TextRank) لتلخيص النصوص الطبية. وأخيرًا يستفيد الحل الثالث من الأنطولوجيات (UMLS) والمصدر (WordNet) اللغوي والمعنوي لتبسيط اللغة الطبية ، و الذي يعتبر مكمل للحلين الأولين.

تقدم هذه الدراسة إسهامًا هامًا في مجال تبسيط النصوص الطبية من خلال اقتراح حلول عملية. يوفر النهج الدلالي المعتمد آفاقًا واعدة لتسهيل فهم واستخدام المعلومات الطبية. يمكن أن يكون تبسيط النصوص الطبية من هذا القبيل له تأثير إيجابي على جودة الرعاية الصحية وتعزيز إمكانية الوصول إلى المعلومات الطبية لجمهور واسع من القراء.

خلاصة القول ، يؤكد هذا البحث على أهمية هذا النهج وتداعياته المحتملة في مجال الصحة.



**الكلمات الرئيسية :** معالجة اللغة الطبيعية، إنتاج النص التلقائي، ملخص النص، تبسيط النصوص، تحليل المعنى.

Abstract

This thesis presents a semantic approach aimed at simplifying medical texts. Three distinct solutions are implemented within this approach. The first solution utilizes the pre-trained model BioGPT to generate medical text. The second solution relies on the TextRank algorithm to summarize medical texts. Lastly, the third solution leverages UMLS ontologies and the lexico-semantic resource WordNet to simplify medical language, complementing the first two solutions.

This research work makes a significant contribution to the field of medical text simplification by proposing concrete solutions. The adopted semantic approach offers promising prospects for facilitating the understanding and utilization of medical information. Such simplification of medical texts can have a positive impact on the quality of healthcare and promote better accessibility to medical information for a wide range of readers. In summary, this thesis underscores the importance of this approach and its potential implications in the healthcare domain.

**Keywords:** Natural Language Processing, Automatic Text Generation, Text Summarization, Text Simplification, Semantic Analysis.

**Remerciements**

Le chemin n'a pas été court et ne devrait pas l'être, mais voici que les années ont passé et ce qui était un rêve d'hier s'est réalisé aujourd'hui el-hamdulilah.

Tout d'abord, je tiens à remercier ALLAH le tout puissant de m'avoir donné la volonté, la patience, la santé et de m'avoir aidé à surmonter toutes les dures épreuves et les moments difficiles pour aboutir à ce travail.

J'adresse tous les plus profonds et plus sincères remerciements à mon encadreur: Mme MEZZI Melyara, pour m'avoir guidé pendant cette année. C'était un honneur de travailler avec elle, je la remercie pour sa disponibilité, son soutien, surtout sa façon de motivation et sa patience durant toute la période du déroulement de ce projet.

Je remercie les membres du jury pour l'honneur et l'intérêt qu'ils m'ont accordé en acceptant d'examiner et d'évaluer ce mémoire.

Mes pensées vont vers mes chers parents. Je tien à remercie celle qui m'a porté inlassablement et qui m'a vu avec son cœur avant ses yeux. Et celui dont je porte son nom avec fierté, qui m'a donné sans limites et continue. Les deux qui ont travaillé d'être que je suis aujourd'hui, et sont toujours l'ombre qui m'abrite tout le temps

Mes remerciements les plus sincères à mon frère et mes sœurs qui m'ont soutenu par leur amour, confiance et qui ont été une aide sur ce chemin. Pour que ses yeux se réjouissent de me voir aujourd'hui.

Je tiens à remercie Brahimi Sarah celle qui je n'ai pas de lien de parenté, mais qui est comme une sœur et un soutien. Elle m'a soutenu dans les moments difficiles et elle a été ma compagne dans mes moments heureux. Sans oublier Laouichi Zineb, la compagne qui a embelli le chemin du succès et qui a été une amie fidèle tout au long de ces 7 années.

Je remercie tous mes ami(e)s, ma famille qui m'ont soutenu tout au long de ce parcours, tous ceux qui lisent ce travail que ce soit pour l'évaluer, le critiquer ou bien pour enrichir leurs connaissances et tous ceux qui ont contribué à la réalisation de ce travail ou qui ont été la cause de ma joie sans oublier toute mes clientes de Candles Corner Store.

Enfin, mes plus sincères remerciements à celui que Dieu choisira pour être mon partenaire, mon soutien et ma belle récompense.

Khaoula

**Remerciements**

Je remercie ALLAH de sa présence constante à mes côtés, sa bienveillance et sa protection m'ont permis de surmonter les difficultés, de trouver la motivation et de persévérer malgré les obstacles.

Ce projet de fin d'étude était un défi, c'est également une occasion de revenir sur les dix-sept dernières années, de réfléchir aux réalisations et aux echecs et d'envisager l'avenir. Aussi stressant que cela ait pu être, cela en valait la peine d'arriver jusqu'ici, pour moi et pour mes proches.

Mes remerciements sont également adressés à ma chère promotrice, Mme MEZZI Melyara, j'ai une immense gratitude envers elle pour les encouragements qu'elle m'a prodigués, son soutien, et même les e-mails qui n'étaient pas liés à ce travail, mais qui ont suscité en moi savoir et confiance. Ils m'ont fait prendre conscience de ma chance. Sans son précieux accompagnement, cette recherche n'aurait jamais pu voir le jour. Avec tous les égards dus à son rang, je lui exprime une gratitude infinie.

Mes chers parents, papa et maman, ainsi que mon frère Chiheb Eddine et ma sœur Meriem. Ces quatre personnes, qui jouent un rôle essentiel dans mon existence, ont été bien plus que présentes tout au long de ce parcours. Leurs prières et leurs paroles réconfortantes m'ont soutenu sans relâche. Je vous remercie du fond du cœur, car sans vous, je n'aurais jamais pu atteindre le niveau où je me trouve aujourd'hui ! Vous êtes une véritable bénédiction dans ma vie, et je suis fière d'appartenir à cette merveilleuse famille. Je suis comblée de vous rendre fiers par l'obtention de ce master.

À mes précieux amis, qui m'ont témoigné leur amitié et leur soutien tout au long de ces années universitaires, je tiens à vous exprimer ma profonde gratitude pour votre sollicitude durant ces moments. Chacun d'entre vous mérite mes remerciements du fond du cœur, de manière toute particulière.

Zineb

## Dédicace

Je dédie ma réussite à mes parents pour leurs sacrifices, leurs encouragements et surtout pour leurs prières. Je leurs souhaite la bonne santé et la longue vie, Incah Allah. À mon frère Mohamed et mes sœurs Rihab et Manar. À tous les membres de mes deux familles GUERBA et RATIAT sans exception, grands et petits. À tous mes amies spécialement Sarah, Zineb et Fella. À Bengouffa F, Chiboub T, Nachef K, Cherguelaine A et Brahim A.

Khaoula

## Dédicace

Suis-je réellement arrivée à ce jour? J'arrive pas à croire mes yeux.

Je dédie ce travail à ma famille que j'adore, vous qui m'avez encouragée à atteindre ce point et à obtenir ce diplôme. "Prière" de respecter mon souhait de m'éloigner complètement de ce domaine hahah!.

À ma deuxième famille, ITCommunity, je vous remercie pour tous ces merveilleux souvenirs que je garderai dans ma mémoire. Je suis reconnaissante de m'avoir accueillie malgré ma nature introvertie et je vous suis reconnaissante de me permettre de devenir un jour votre animatrice. Il y a plusieurs personnes à mentionner individuellement, mais sachez que je vous aime bien tous.

GDG Community, où j'ai vu le monde s'élargir et grandir, une famille remplie d'énergie et de sourires, j'ai vécu des moments spéciaux avec vous, et je vous remercie d'avoir été si généreux dès notre première rencontre. Une mention spéciale aux membres de Let's Discover team 4 de Let's manage (best presentation en GIP 2023).

En particulier, je dédie cette thèse à une personne trés chère à mon cœur, ma deuxième sœur, ma chamelle, my soulmate. Que je remercie pour ses encouragements énormement et son soutien moral, avec la quelle j'ai partagé les bons et les mauvais moments, je suis trés fière d'elle pour sa réussite, my MASTERCHEF (pas besoin de le prouver), à la plus douce, la plus adorable, la plus tout B.Sarah.

À mes chers amis: Dj.Mohamed, M.Wassim, A.Ayoub, B.Fady, ma chingu El.Sihem, my mentor E.Hadjer, ma O.Fella with an a, Z.Sara, ainsi que mon binôme G.Khaoula. Je tiens à vous exprimer ma gratitude pour l'amitié, le respect, les encouragements, les beaux messages et l'aide que vous m'avez apportés, je suis sincèrement heureuse de vous avoir durant mon parcours.

ATTACK ON TITAN, je n'aurais jamais imaginé pouvoir aimer autant un anime. Pour moi, il a été présent tout au long de mon parcours, créant de beaux souvenirs. Nous approchons de la fin, avec seulement un film restant. EREHH, "Jiyuu da"!

Je ne peux pas terminer ces dédicaces sans mentionner les épisodes amusants de Going Seventeen: Mafia game. Merci de m'en procurer du soulagement, de me détendre et de me faire rire. la présence de GoSe a été un véritable remède contre le stress.

Je vous aime tous, et je suis enfin libre de dormir d'un sommeil profond, d'un sommeil de plomb.

Zyneep

# Table des Matières

# Liste des Figures

# Liste des Tableaux

# Liste des abréviations

| | |
|---|---|
| TALN | Traitement Automatique du Langage Naturel. |
| NER | Named Entity Recognition. |
| RNN | Recurrent Neural Networks. |
| TAN | Traduction Automatique Neuronale. |
| DRESS | Deep Reinforcement Sentence Simplification. |
| MST | Modèle de Simplification de Texte. |
| AEM | Algorithme d'Espérance-Maximisation. |
| TXT | Text. |
| JSON | JavaScript Object Notation. |
| GEM | Generation Evaluation Metrics. |
| NIH | National Institutes of Health. |
| CNN | Convolutional Neural Network. |
| BERT | Bidirectional Encoder Representations from Transformers. |
| GPT | Generative Pre-trained Transformer. |
| SDoH | Social Determinants of Health. |
| UMLS | Unified Medical Language System. |
| SNOMEDCTUS | Systematized NOmenclature of MEDicine-Clinical Terms United States. |
| HPO | Human Phenotype Ontology. |
| MeSH | Medical Subject Headings. |
| CUI | Unique Concept Identifier. |
| WordNet | Word Network. |
| ROUGE | Recall-Oriented Understudy for Gisting Evaluation. |

# Introduction Générale

## Contexte Global

La diffusion adéquate de l'information médicale revêt une importance primordiale en vue d'assurer la compréhension et l'accessibilité des connaissances médicales, notamment pour un vaste public. Malheureusement, les textes médicaux se caractérisent souvent par leur complexité et leur difficulté à appréhender, ce qui représente un obstacle majeur pour de nombreux individus. Dans cette perspective, la simplification des textes médicaux apparaît comme une solution prometteuse en vue de faciliter l'accès et la compréhension de ces informations.

## Problématique

La problématique soulevée dans ce travail de recherche réside dans la complexité des textes médicaux en anglais et les difficultés qu'elle engendre pour la compréhension et l'accès à ces informations par un large public. De nombreux individus, tels que les patients, les aidants, ou même les professionnels de la santé non spécialisés, peuvent être confrontés à des obstacles lorsqu'ils essaient de lire et de comprendre ces textes. Ainsi, il devient essentiel de développer des méthodes et des outils permettant de simplifier ces textes médicaux sans compromettre leur contenu informatif et leur précision. Quel impact pourrait avoir une accessibilité accrue aux informations médicales sur la perception globale du domaine médical? En permettant à n'importe quelle personne de comprendre le monde médical, quelles conséquences cela pourrait-il entraîner sur la réduction des dommages potentiels associés à la méconnaissance de ce domaine?

## Objectifs de la Recherche

Notre recherche vise à exploiter le potentiel du traitement du langage naturel dans le domaine médical afin de concevoir une application dédiée à la génération et à la simplification des textes médicaux, en intégrant une analyse sémantique. Nous nous appuierons sur les sous-domaines spécialisés du traitement automatique du langage naturel pour le traitement des données textuelles. Les techniques associées nous permettront de prétraiter et de

nettoyer nos données en vue de leur utilisation dans l'entraînement ou dans la génération des textes.

## Organisation du Mémoire

Dans cette thèse, nous entamons tout d'abord une analyse approfondie de la simplification des textes médicaux, en examinant les travaux existants dans ce domaine.

Par la suite, nous décrivons en détail notre conception de recherche, en présentant nos trois solutions proposées et en décrivant les différentes étapes de traitement des données. Nous mettons également en évidence les différents traitements de texte ainsi que l'enrichissement sémantique visant à améliorer la simplification du texte généré.

Dans le troisième et dernier chapitre, nous exposons la mise en œuvre concrète de nos trois approches proposées, accompagnée de leur évaluation et de la vérification des résultats à l'aide de cas de test pertinents. De plus, nous présentons notre interface d'application qui intègre les trois solutions développées.

Enfin, nous concluons ce travail par une conclusion générale, mettant en évidence les principaux résultats et contributions de notre étude. Nous abordons également les perspectives futures envisagées pour poursuivre les avancées dans le domaine de la simplification des textes médicaux.

# Chapitre 1

# État de l'Art

## 1.1 Introduction

Avec l'évolution technologique du Web, la documentation en matière de santé a connu une croissance exponentielle, de plus en plus accessible à tous, y compris aux patients, qui peuvent accéder à une mine d'informations sur la santé. Cependant, cette richesse d'informations médicales ne garantit pas automatiquement sa bonne compréhension par le public cible, en particulier les non-spécialistes ayant peu ou pas de connaissances médicales [1].

Ce chapitre sur la simplification automatique des textes aborde diverses notions et techniques utilisées dans le domaine. Au fil des recherches, plusieurs méthodes ont été élaborées pour simplifier les textes médicaux. Parmi ces méthodes, on retrouve la Normalisation des Abréviations Médicales, la Désambiguïsation Lexicale, l'Apprentissage Automatique dans le Domaine Médical, etc, de même que les Techniques de sémantification des textes, telles que la morphologie, la syntaxe, la lexicale, la sémantique et la pragmatique. Ces différentes approches offrent des moyens efficaces pour simplifier et faciliter la compréhension des textes médicaux.

En ce qui suit, nous nous intéresserons de près à la simplification automatique des textes médicaux, ses approches, ses techniques et à ses travaux connexes.

## 1.2 Simplification Automatique des Textes Médicaux

Le vocabulaire médical, avec ses termes techniques et ses acronymes mystérieux, peut sembler à bien des égards comme une langue étrangère pour les personnes non initiées. Cette difficulté de compréhension peut rendre l'accès à l'information médicale difficile, voire impossible, pour ceux qui cherchent à en savoir plus sur leur santé. Malheureusement, cette barrière linguistique peut créer une distance et une incompréhension entre les professionnels de la santé et les patients, ce qui peut avoir des conséquences néfastes sur la qualité des soins et le suivi des traitements. En somme, l'utilisation d'un langage trop technique

peut se révéler contre-productive et compliquer inutilement la communication entre les professionnels de la santé et leur public [2].

Le présent texte constitue un exemple de document médical renfermant des termes relativement complexes à appréhender:

*"La pathologie hépatique biliaire intrahépatique est associée à une diminution de l'activité de la phospholipide flippase, conduisant à une accumulation de phosphatidylcholine dans les hépatocytes."*[1].

Afin de rendre les informations cruciales plus accessibles, diverses méthodes et technique de simplification des textes médicaux ont été élaborées. Ces approches exploitent les avancées dans le domaine du traitement automatique du langage naturel (TALN) et de la recherche en biomédecine, dans le but de diminuer la complexité des textes médicaux tout en préservant leur précision et leur pertinence.

### 1.2.1 Reconnaissance des Entités Nommées Médicales

Parmi les techniques avancées figurent la reconnaissance d'entités nommées (NER), est une technique clé utilisée dans le domaine médical permettant d'identifier et extraire les entités spécifiques présentes dans le texte. Cette technique vise à repérer et à classer automatiquement des entités telles que les noms, les noms des maladies, les procédures médicales, les symptômes, les dates, et autres, présentes dans le texte. Cependant, la réussite de cette technique dépend étroitement des entités spécifiques incluses dans le modèle utilisé[3].

Afin d'illustrer plus clairement le concept, nous proposons un exemple concret dans la figure1.1 suivante en utilisant le modèle "ukkendane/bert-medical-ner" sur Hugging Face[2].

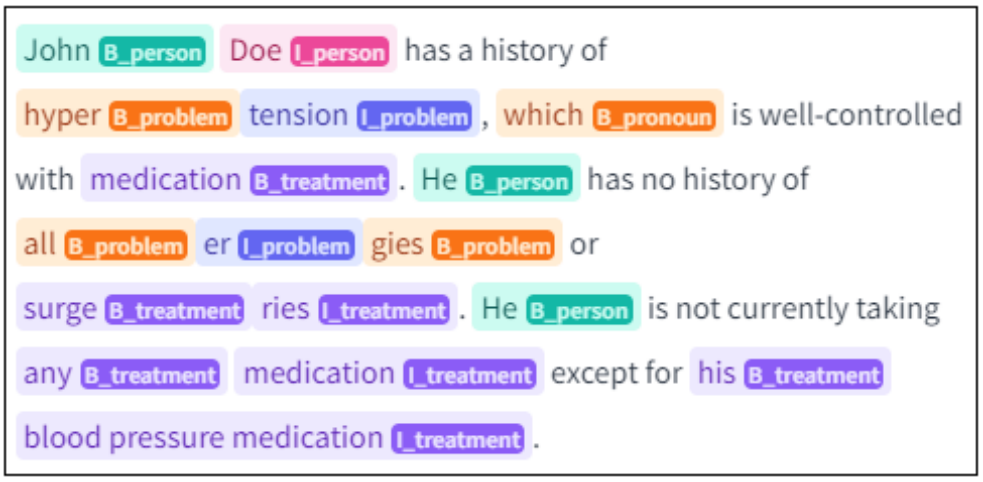


Figure 1.1: Exemple de la reconnaissance des entités médicales[3].

[1]http://participants-area.bioasq.org/general-information/Task8b.
[23]https://huggingface.co/ukkendane/bert-medical-ner.

### 1.2.2 Normalisation des Abréviations Médicales

Une autre technique, la normalisation des abréviations médicales constitue une approche visant à convertir les abréviations en termes médicaux complets et compréhensibles. Dans le domaine médical, les abréviations sont couramment utilisées pour représenter des termes techniques et des expressions complexes. Toutefois, ces abréviations peuvent engendrer de la confusion, notamment pour ceux qui ne sont pas familiers avec leur signification. L'objectif de la normalisation des abréviations est d'identifier et de substituer ces abréviations par leur forme complète ou leur signification équivalente. Cette pratique vise à améliorer la clarté et l'accessibilité du texte médical, facilitant ainsi la compréhension et l'interprétation des informations fournies[4].

L'illustration1.2 ci-dessous montre un exemple des variantes lexicales les plus fréquentes comportant deux sens ou plus, classées selon le type de distribution, utilisées dans le cadre de l'étape de normalisation des abréviations médicales.

| Short form term | Total count | Senses according to concept unique identifiers | Distribution of senses |
|---|---|---|---|
| 'pt' | 137 | C0030705: Patients | 89 % |
| | | C0949766: Physical therapy procedure | 4 % |
| | | C0086835: Structure of the posterior tibial artery | 4 % |
| | | 3 more senses | 8 % |
| 'ct' | 82 | C0040405: X-Ray computed tomography | 95 % |
| | | C1274037: Cardiothoracic surgery | 2 % |
| | | C0008034: Thoracic drain | 2 % |
| | | 1 more sense | 1 % |
| 'm' | 62 | C0024554: Male gender | 81 % |
| | | C0018808: Heart murmur | 16 % |
| | | C0026591: Mother | 2 % |
| | | 1 more sense | 2 % |

Figure 1.2: Exemple des variantes lexicales utilisées dans la normalisation des abréviations médicales[4].

### 1.2.3 Désambiguïsation Lexicale

D'autres techniques, telles que la désambiguïsation lexicale qui vise à éclaircir le sens des termes ambigus[5], par exemple, l'image1.3 ci-dessous représente un schéma de principe de la désambiguïsation lexicale supervisée, décrivant de manière visuelle les différentes étapes et processus impliqués dans cette approche.

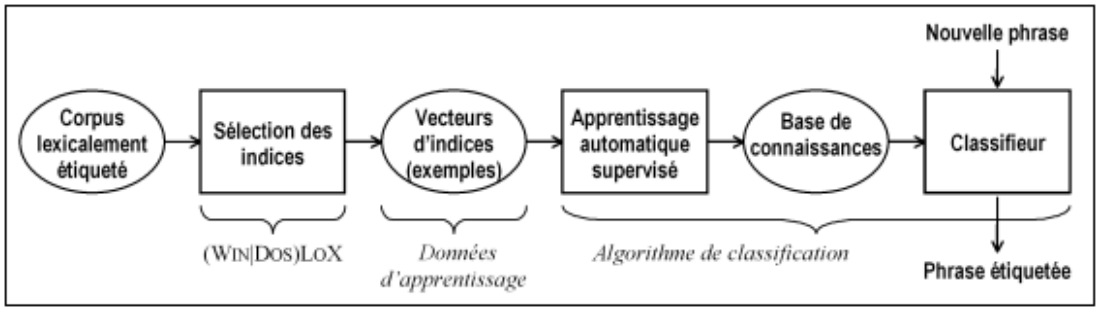


Figure 1.3: Schéma pour la désambiguïsation lexicale supervisée[5].

### 1.2.4 Apprentissage Automatique dans le Domaine Médical

Le domaine de la simplification des textes médicaux explore également des approches plus avancées. Parmi celles-ci, on trouve l'utilisation de techniques d'apprentissage automatique et de modèles neuronaux tels que les réseaux de neurones récurrents (RNN[4]) et les Transformers[6], qui permettent de saisir les relations et les structures complexes présentes dans les phrases médicales. Ces modèles peuvent être entraînés sur de vastes corpus de textes médicaux afin d'apprendre les motifs linguistiques et de générer des reformulations simplifiées.

Le schéma1.4 suivant est exemple qui montre le détail d'une couche récurrente simple. Les $w_j^i$ et les $y_j^i$ désignent respectivement les entrées et les sorties de la couche à l'instant $t$.

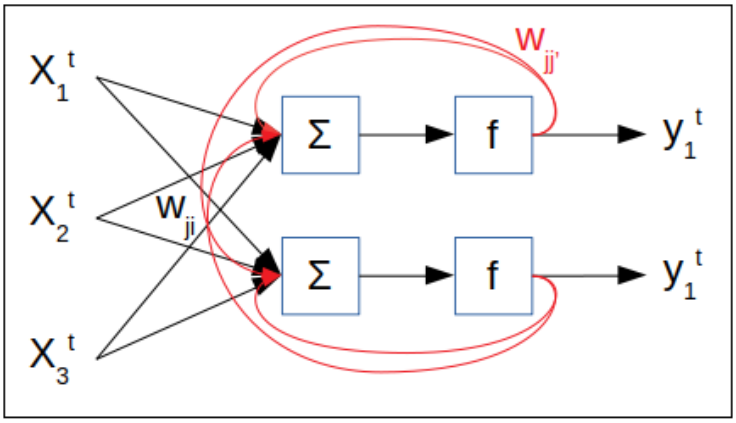


Figure 1.4: Exemple d'une couche RNN simple à trois entrée et deux sorties. Les connexions récurrentes sont notées en rouge[7].

[4]Les réseaux de neurones récurrents (RNN) sont un type d'architecture de réseau de neurones qui permet de traiter des données séquentielles ou temporelles. Contrairement aux réseaux de neurones traditionnels, les RNN ont des connexions récurrentes qui leur permettent de conserver une mémoire interne et d'exploiter les informations contextuelles des données précédentes.

La présence de poids $w_j^i$ reliant les entrées à la sortie, et la présence de poids $r_{jj'}$ entre la sortie et l'entrée de la couche, qui sont les fameuses connexions récurrentes (en rouge). Le calcul de la sortie d'une couche de neurones peut donc se faire par l'équation:

$$y_j^t = f(\sum_i W_{ji} x_i^t + \sum_{j'} r_{ii'} y_i^{t-1})$$

,où le deuxième terme modélise la récurrence du réseau.

### 1.2.5 Transcription Automatique de la Parole dans le Domaine Medical

Il est à souligner que la simplification des textes médicaux transcende le domaine de la langue écrite. Des méthodes de simplification sont également employées dans la transcription automatique de la parole afin de convertir les enregistrements audio des consultations médicales en textes écrits plus clairs et compréhensibles[8].

L'image1.5 ci-dessous présente un exemple d'un schéma général pour la transcription automatique de la parole.

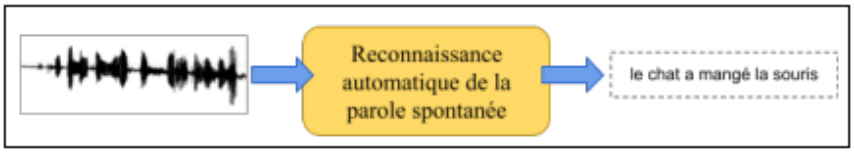


Figure 1.5: Schéma général pour la transcription automatique de la parole[9].

Il convient de noter que la simplification des textes médicaux représente un défi complexe et pluridimensionnel. Les chercheurs s'efforcent constamment de développer de nouvelles approches et techniques afin d'améliorer la qualité et l'efficacité de la simplification, tout en garantissant l'exactitude et la précision des informations médicales.

### 1.2.6 Techniques de Sémantification des Textes

Dans le domaine de la sémantification des textes, plusieurs techniques sont employées pour extraire et représenter le sens des mots et des phrases. Ces techniques peuvent être classées en trois catégories principales: celles basées sur le dictionnaire, celles basées sur les ressources sémantiques et celles basées sur la pragmatique, notamment les ontologies[10].

Dans cette section, nous allons explorer ces différentes approches en les articulant autour des niveaux d'analyse en Traitement Automatique du Langage: morphologique, syntaxique, lexicale, sémantique et pragmatique[11].

Pour la sémantification des textes, il est essentiel de mettre en place un pipeline efficace qui permet d'analyser et de traiter les différentes dimensions linguistiques afin de capturer

le sens des termes et des phrases de manière précise. La figure1.6 suivante est un exemple de pipeline idéal pour la sémantification des textes.

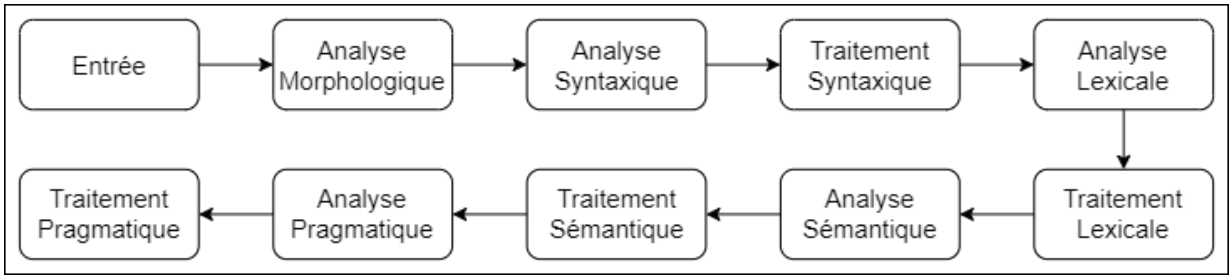


Figure 1.6: Pipeline idéal pour la sémantification des textes.

### Analyse Morphologique

L'analyse morphologique est une étape essentielle dans le traitement automatique du langage naturel qui se concentre sur l'étude des structures morphologiques des mots. Elle permet de segmenter les mots en unités plus petites appelées morphèmes et d'analyser leur formation et leur fonction grammaticale. Cette analyse morphologique joue un rôle crucial dans de nombreux domaines, tels que la lemmatisation, la dérivation, la flexion verbale, la désambiguïsation morphologique, etc[12].

Pour réaliser l'analyse morphologique, différentes approches peuvent être utilisées, allant des méthodes basées sur des règles linguistiques à celles basées sur l'apprentissage automatique. Les ressources lexicales telles que les dictionnaires et les lexiques morphologiques sont également utilisées pour extraire des informations sur les formes et les propriétés morphologiques des mots[13].

Pour mieux comprendre l'analyse morphologique, prenons l'exemple du mot "inattendu" comme le montre le tableau1.1 suivant. L'analyse morphologique de ce mot révèle qu'il est composé du préfixe "in-" (qui exprime la négation) et du radical "attendu". En segmentant le mot en morphèmes, nous pouvons observer comment les éléments constitutifs contribuent à la formation et au sens global du mot.

| **Mot** | **Préfixe** | **Radical** |
|---|---|---|
| inattendu | in | attendu |

Table 1.1: Exemple d'une analyse morphologique.

### Analyse Syntaxique

L'analyse syntaxique est une étape cruciale dans le traitement automatique du langage naturel. Elle permet d'identifier les différents constituants syntaxiques et les relations de dépendance entre eux. Cette analyse revêt une importance fondamentale car elle fournit une représentation structurelle qui facilite la compréhension et l'interprétation du

texte. Plusieurs approches et techniques du traitement syntaxique ont été développées pour l'analyse syntaxique, chacune avec ses propres caractéristiques et avantages. Parmi ces approches, on retrouve l'analyse syntaxique basée sur des règles, l'analyse syntaxique statistique et l'analyse syntaxique basée sur des modèles d'apprentissage automatique. Ces méthodes peuvent être appliquées à différents niveaux de granularité, allant de l'analyse des mots individuels jusqu'à l'analyse complète des phrases[14].

Pour mieux comprendre l'analyse syntaxique avec le traitement syntaxique, prenons l'exemple1.2 suivant, l'analyse syntaxique nous permet de comprendre la structure grammaticale de la phrase et les rôles des différents mots qui la composent.

| Phrase | Le chat mange une souris. |
|---|---|
| **chat** | nom commun et occupe le rôle de sujet. |
| **mange** | verbe conjugué à la troisième personne du singulier. |
| **une** | déterminant indéfini. |
| **souris** | nom commun et occupe le rôle de complément d'objet direct. |

Table 1.2: Exemple d'un traitement syntaxique.

### Analyse Lexicale

L'analyse lexicale, également connue sous le nom d'analyse lexicographique, est une étape fondamentale dans le traitement automatique du langage naturel (TAL). Elle vise à extraire et à analyser les informations lexicales contenues dans un texte afin d'obtenir une compréhension approfondie de son contenu lexical. Cette analyse joue un rôle crucial dans de nombreux domaines d'application, tels que la recherche d'information, la traduction automatique, l'analyse de sentiment, etc[15].

Le traitement lexical, qui suit l'analyse lexicale, concerne les manipulations et les opérations effectuées sur les unités lexicales extraites. Cela peut inclure des tâches telles que la normalisation des mots (par exemple, la réduction des mots à leur forme canonique) ou la lemmatisation (c'est-à-dire trouver le lemme d'un mot)[16].

Les techniques basées sur le dictionnaire sont largement utilisées dans la sémantification des textes. Ces techniques sont utilisées pour l'analyse lexicale, et également sont utilisées pour enrichir la compréhension sémantique d'un texte. Elles reposent sur l'exploitation de lexiques et de dictionnaires pour attribuer des significations aux mots. Ces ressources fournissent des informations sur les sens des mots, leurs synonymes, leurs relations sémantiques et d'autres caractéristiques lexicales. L'analyse lexicale permet ainsi d'enrichir la compréhension du texte en associant à chaque terme des informations sémantiques provenant du dictionnaire[17].

Pour illustrer ces concepts, prenons l'exemple suivant: dans un corpus médical, nous pouvons rencontrer le terme "maladie". L'analyse lexicale nous permettra d'identifier ce terme et de le désambiguïser en fonction du contexte spécifique. Ensuite, lors du traitement lexical, nous pouvons appliquer des opérations de normalisation pour réduire le terme à sa forme canonique, par exemple en le transformant en "malade". Ces opérations de

traitement lexical sont essentielles pour garantir la cohérence et la précision des résultats obtenus lors du traitement du texte.

### Analyse Sémantique

L'approche sémantique est une méthode de traitement de texte sophistiquée qui se concentre sur la compréhension du sens et de la signification des mots et des phrases, plutôt que sur leur simple structure grammaticale. Elle repose sur une solide connaissance linguistique et des modèles de compréhension du langage naturel pour traiter le texte de manière plus efficace [18].

Dans le domaine de la simplification des textes médicaux, l'approche sémantique offre une approche novatrice pour identifier les concepts clés et les relations entre eux afin de simplifier le texte sans en altérer la signification médicale. Cette méthode permet de traduire des termes techniques complexes en un langage plus accessible pour les patients et le grand public, tout en préservant l'exactitude et la précision des informations médicales. L'approche sémantique constitue ainsi un puissant outil pour améliorer la compréhension des textes médicaux, permettant aux non-experts de mieux appréhender les informations tout en maintenant un haut niveau de qualité médicale[19].

En somme, l'approche sémantique représente une avancée significative dans la simplification des textes médicaux, offrant une solution innovante et en combinant les connaissances linguistiques avec des technologies de pointe, cette méthode offre une réponse efficace aux défis posés par la complexité de la terminologie médicale, permettant ainsi de faciliter l'accès à l'information médicale pour tous les publics [20].

La Figure1.7 illustre un schéma sémantique qui permet de mieux comprendre les relations entre les différents concepts utilisés.

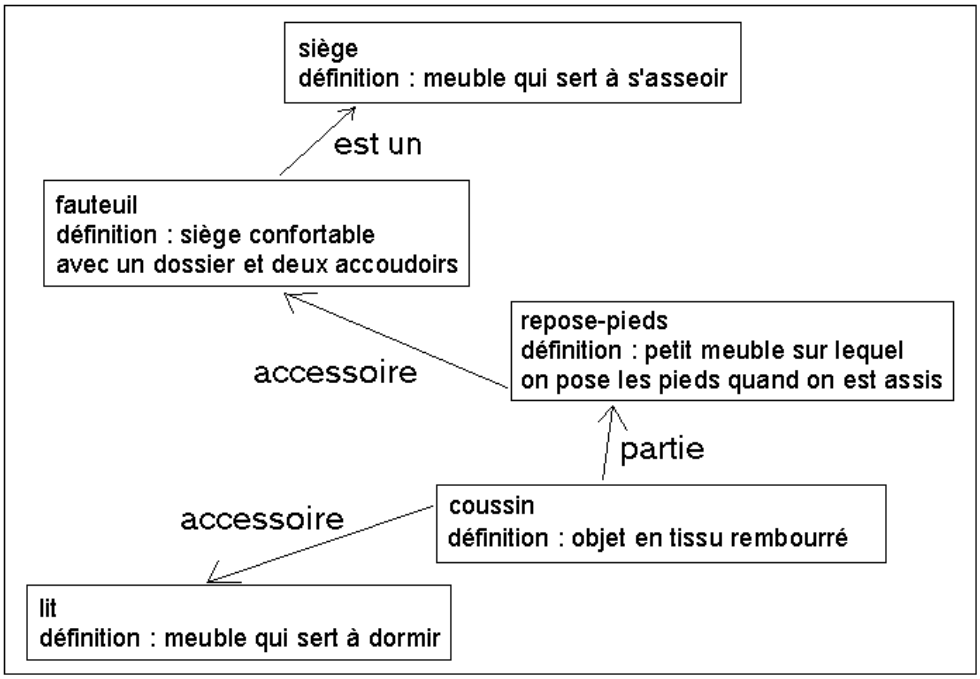


Figure 1.7: Réseau Sémantique[5].

### Analyse Pragmatique

L'analyse pragmatique est une branche essentielle de la linguistique qui vise à étudier le rôle du contexte et des intentions communicatives dans la compréhension du sens d'un énoncé. Elle se concentre sur la manière dont les locuteurs utilisent et interprètent le langage dans des situations de communication réelles. L'analyse pragmatique permet de dépasser la signification littérale des mots et des phrases pour saisir leur sens réel dans un contexte donné. Une fois que les étapes d'analyse morphologique, syntaxique et sémantique ont été réalisées, l'analyse pragmatique entre en jeu pour affiner l'interprétation du texte[21].

Dans le cadre de l'analyse pragmatique, le traitement pragmatique est une étape essentielle qui consiste à appliquer les connaissances pragmatiques acquises pour une interprétation plus précise du texte. Cela peut inclure l'ajustement de l'interprétation en fonction du contexte, la déduction d'informations implicites ou non explicitées, et la prise en compte des aspects socioculturels qui influencent la signification[22].

Pour mieux illustrer ces concepts, prenons l'exemple suivant: supposons qu'un locuteur dise: "Il fait chaud ici." L'analyse pragmatique permet de comprendre que le sens réel de cette phrase dépend du contexte dans lequel elle est prononcée. Si elle est dite dans une pièce fermée et que les gens transpirent, nous pouvons inférer que le locuteur veut exprimer son inconfort et son désir d'aérer la pièce. En revanche, si cette phrase est prononcée dans un cadre informel où les gens portent des vêtements légers et profitent du soleil, le sens peut être simplement descriptif.

[5]https://cui.unige.ch/isi/cours

La pragmatique joue également un rôle crucial dans la sémantification des textes, notamment grâce à l'utilisation des ontologies. Les ontologies fournissent des structures formelles pour représenter les connaissances dans un domaine spécifique. Elles permettent de modéliser les concepts, les relations et les contraintes sémantiques, facilitant ainsi l'interprétation et la manipulation du sens des termes dans le texte. Les ontologies offrent une vision plus large de la signification en prenant en compte le contexte et les inférences logiques[23].

En conclusion, la sémantification des textes repose sur différentes techniques, telles que celles basées sur le dictionnaire, les ressources sémantiques et la pragmatique. En combinant les niveaux d'analyse en TAL, il est possible d'obtenir une compréhension approfondie du sens des termes et des phrases dans un texte donné.

Pour notre sémantification nous avons adopté dans la section suivante deux nouvelles pustes qui sont la génération et le résumé des textes.

On s'intéresse à la génération de texte pour augmenter et enrichir les données, car les utilisateurs n'ont pas une idée très claire et précise de ce qu'ils font rechercher.La deuxième hypothèse qui nous intéresse est le résumé de texte où le contenu médicale qu'on peut trouver sur le web est très long et complexe pour être adéquatement analyser pour un simple humain.

.

## 1.3 La Génération de Texte

La génération de texte en traitement automatique du langage naturel (TALN) vise à créer automatiquement du texte cohérent et compréhensible par les machines. Elle cherche à produire du texte qui semble rédigé par un être humain, en respectant les règles grammaticales, la cohérence sémantique et le style approprié[24].

La génération de texte en TALN est utilisée dans divers domaines tels que la rédaction automatique, les résumés, les dialogues et les chatbots. Cependant, elle présente des défis liés à la cohérence, à la compréhension du contexte et à la prévention des biais indésirables dans le texte généré[25].

Pour cette tâche, différentes méthodes sont disponibles, en voici quelques-unes:

- **Modèles basés sur les règles:** Ces modèles utilisent des règles spécifiques définies par les développeurs pour générer du texte. Ils peuvent être utiles dans des domaines spécifiques où les structures de phrases sont régies par des règles claires[26].
- **Modèles de traduction automatique:** Ces modèles sont conçus pour traduire automatiquement du texte d'une langue source à une langue cible. Ils peuvent également être utilisés pour générer du texte en utilisant l'entrée dans une langue source et en produisant une traduction dans la même langue[27].

- **Modèles de langue probabilistes:** Les modèles de langue probabilistes sont des modèles qui utilisent des probabilités pour prédire la probabilité d'un mot suivant dans une séquence de mots. Ils sont souvent utilisés dans le traitement automatique du langage naturel pour la reconnaissance de la parole, la traduction automatique, la correction orthographique, la suggestion de mots, etc. Par exemple, dans un modèle de langage en français, la probabilité de la séquence "tous les matins je bois du café" sera supérieure à la probabilité de la séquence "du café je tous les matins bois". Les modèles de langue probabilistes sont basés sur les chaînes de Markov, où la probabilité d'une séquence de mots est le produit des probabilités de chaque mot sachant les mots précédents[28].

- **Modèles basés sur les méta heuristiques:** Les modèles basés sur les méta heuristiques sont des modèles qui utilisent des algorithmes d'optimisation pour résoudre des problèmes complexes en générant du texte. Ces modèles sont souvent utilisés pour la génération de texte créatif, comme la poésie, la musique, les histoires, etc. Les méta heuristiques sont des algorithmes qui cherchent à trouver des solutions optimales à des problèmes en explorant l'espace des solutions possibles. Les modèles basés sur les méta heuristiques utilisent des techniques telles que la recherche tabou, la recherche locale, l'algorithme génétique, etc. pour générer du texte[29].

- **Modèles neuronaux:** Les modèles neuronaux sont des modèles qui utilisent des réseaux de neurones artificiels pour générer du texte. Ces modèles sont souvent utilisés pour la traduction automatique, la génération de texte, la réponse aux questions, etc. Les réseaux de neurones sont des modèles mathématiques qui simulent le fonctionnement du cerveau humain en utilisant des couches de neurones interconnectés. Les modèles neuronaux sont capables de générer du nouveau contenu à partir de la combinaison de données précédemment analysées et peuvent réaliser une large gamme de tâches de traitement de langage, comme la génération de textes, la traduction ou la classification de contenus. Les modèles neuronaux les plus couramment utilisés sont les réseaux de neurones récurrents (RNN), les réseaux de neurones à convolution (CNN) et les réseaux de neurones transformer[30].

- **Modèles de langage pré-entraînés:** Ces modèles, comme BERT et ses variantes, GPT et ses variantes, etc, sont entraînés sur de vastes corpus de texte et peuvent générer du texte cohérent et contextuellement approprié en fonction de l'entrée donnée[31].

## 1.4 Le Résumé de Texte

Le résumé de texte consiste à extraire les informations clés et les points importants d'un texte source afin de produire un résumé concis. L'objectif est de fournir une version condensée du texte qui conserve les informations essentielles[32].

Il existe plusieurs approches pour créer un résumé, notamment:

- **L'approche abstraite:** où le résumé est généré en utilisant des techniques de génération de texte pour créer de nouvelles phrases qui n'apparaissent pas dans le texte source[33].
- **L'approche extractive:** où des phrases ou des passages sont sélectionnés directement à partir du texte source pour créer le résumé[33].
- **L'approche hybride:** qui combine des éléments des approches abstraites et extractives pour créer un résumé plus fluide et précis[34].

## 1.5 Notions Connexes

Dans cette section, nous allons résumé quelques une des notions connexes élémentaires utilisés au cours de notre travail pour faciliter la compréhension de notre solution proposée, tel que les notions TAL qui résument les téchniques de traitement automatique de langage, les notions neuronales qui représent les modèles de langage pré-entraîné et leurs architectures et les notions des ontologies médicales.

### 1.5.1 Notions TAL

- **La tokenisation:** est une technique fondamentale de prétraitement de texte dans le traitement du langage naturel. Cela implique la division systématique d'un texte en unités individuelles appelées jetons.

  L'objectif principal de la tokenisation est de faciliter l'analyse, la manipulation et le traitement des données textuelles dans diverses applications NLP. En segmentant le texte en jetons discrets, il devient plus facile pour les ordinateurs de comprendre et de travailler avec le langage. Ce processus est essentiel pour un large éventail de tâches, notamment la traduction automatique, l'analyse des sentiments, la classification de texte et la reconnaissance d'entités nommées.

- **La Lemmatisation:** *"La lemmatisation fait généralement référence à la réalisation correcte des choses à l'aide d'une analyse lexicale et morphologique des mots, visant normalement à supprimer uniquement les terminaisons flexionnelles et à retourner la forme de base ou lexicale d'un mot, connue sous le nom de lemme."*[6]

  La lemmatisation est une technique de traitement du langage naturel utilisée pour normaliser les mots en linguistique. L'objectif de cette technique est de réduire les mots à leur forme de base ou de racine, ce qui facilite l'analyse et la comparaison de texte. Elle utilise des règles grammaticales et des dictionnaires pour réduire les mots à leur forme de base ou la forme canonique, appelée "lemme". La figure1.8 suivante illustre le fonctionnement de l'algorithme de lemmatisation.

[6]https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

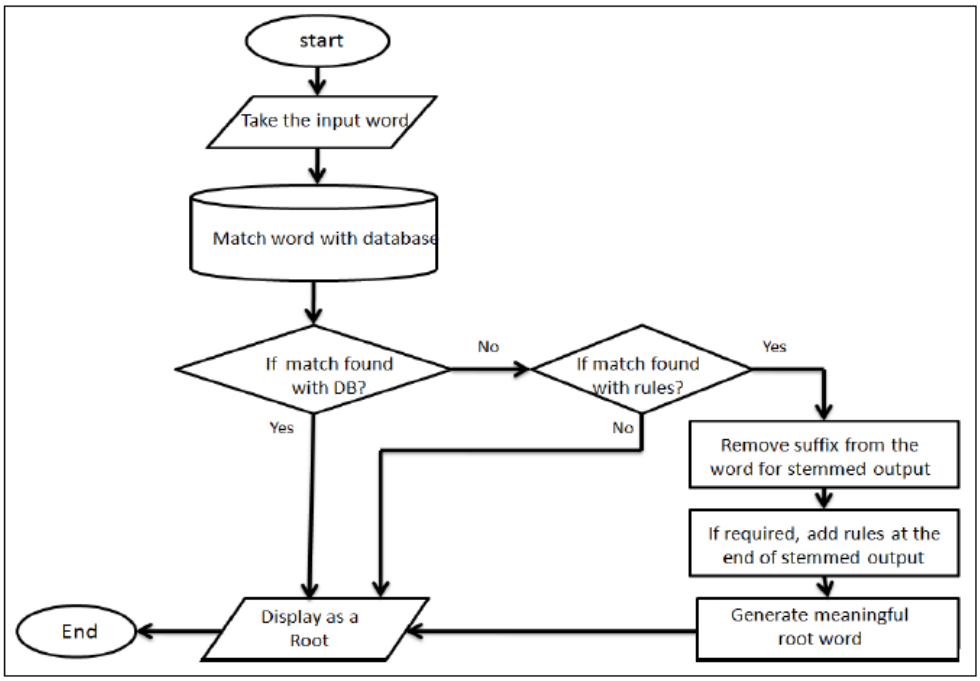


Figure 1.8: Fonctionnement de l'algorithme de lemmatisation[35].

- **Extraction des Entités Nommées:** Dans le domaine médical, l'extraction des entités nommées[36] consiste à identifier et à extraire des informations spécifiques liées au domaine médical à partir d'un texte. Cela peut inclure des entités telles que les noms de maladies, de médicaments, de symptômes, de procédures médicales, de professionnels de la santé, etc.

  L'objectif de cette téchnique est de repérer et de structurer ces entités pour faciliter l'analyse, la recherche d'informations et l'extraction de connaissances médicales à partir des textes.

  Ce processus de reconnaissance d'entités automatisé facilite considérablement l'analyse du texte et fournit des informations précieuses pour une meilleure compréhension du contenu.

  Le diagramme1.9 ci-dessous présente le pipeline utilisé dans le modéle en-ner-bionlp13cg-md pour extraire les entités nommées.

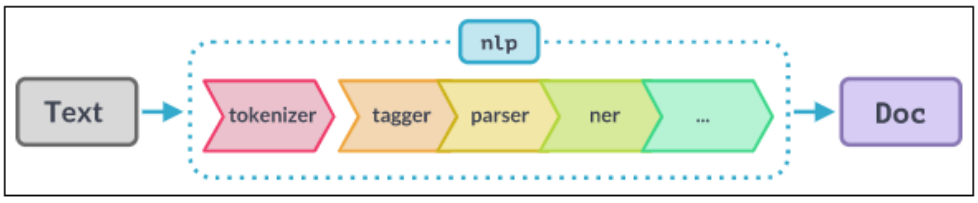


Figure 1.9: Schéma du pipeline du modéle en-ner-bionlp13cg-md[7].

[7]https://spacy.io/api

### 1.5.2 Notions sur les Architectures Neuronales

- **BioBERT:**[8](Biomedical Bidirectional Encoder Representations from Transformers), est un modèle de langage pré-entraîné spécifiquement conçu pour le domaine biomédical. Il repose sur l'architecture des Transformers et a été adapté à partir du modèle BERT (Bidirectional Encoder Representations from Transformers). BioBERT est entraîné sur de vastes corpus de données biomédicales, tels que des articles scientifiques et des publications médicales, afin de capturer les spécificités terminologiques propres à ce domaine. Grâce à son architecture et à son entraînement sur des données biomédicales, BioBERT offre des représentations contextuelles riches pour le traitement du langage biomédical, permettant ainsi de réaliser des tâches de classification, d'extraction d'informations et de génération de texte dans ce domaine[37].

  La figure ci-dessous1.10 est un exemple qui montre l'architecture de BioBERT pour la classification de textes SDoH (Social Determinants of Health). Cette architecture est une implémentation d'un encodeur de transformateur bidirectionnel multicouche. Les entrées de BioBERT sont des séquences de tokens de texte et les sorties sont des représentations vectorielles de chaque token. Ces représentations vectorielles peuvent ensuite être utilisées pour effectuer des tâches de traitement de langage naturel, telles que la classification de textes ou la génération de textes[38].

[8]https://github.com/dmis-lab/biobert.

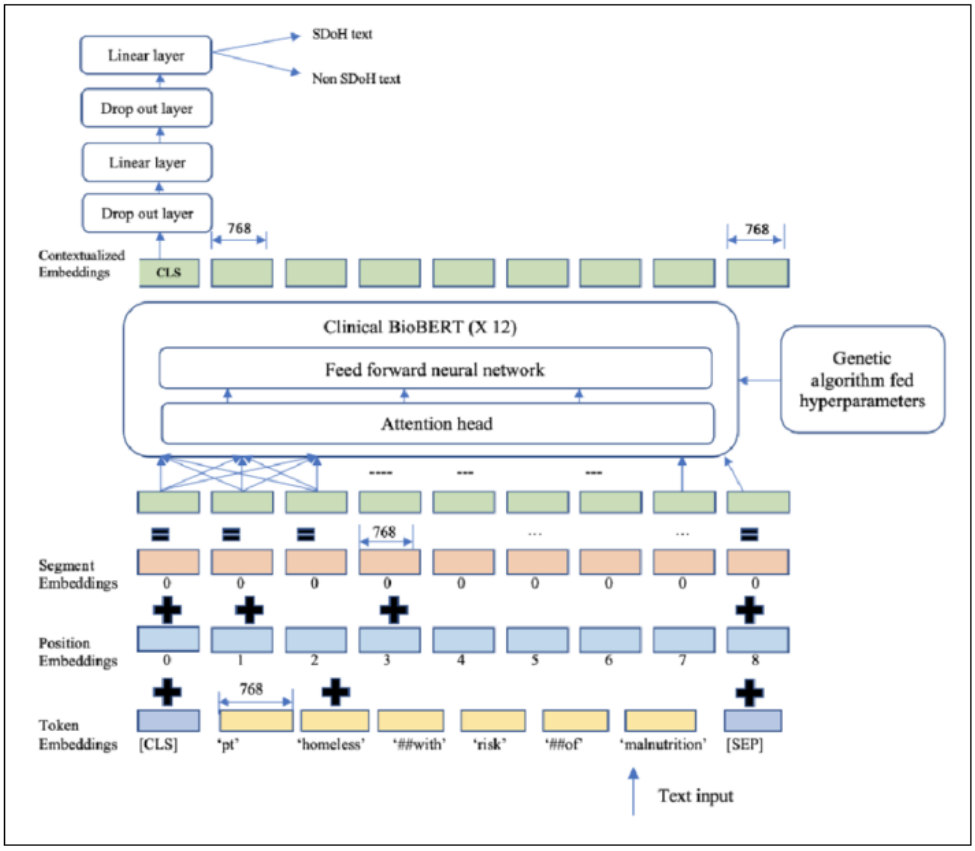


Figure 1.10: Architecture de BioBERT pour la classification de textes SDoH[45].

- **BioGPT:**[9](Transformateur pré-entraîné génératif pour la génération et l'exploration de textes biomédicaux), développé par Microsoft, est un modèle de langage pré-entraîné spécifiquement conçu pour le traitement du langage naturel (TALN) dans le domaine biomédical. Il est basé sur l'architecture GPT (Generative Pre-trained Transformer) et a été entraîné sur un large corpus de littérature biomédicale, comprenant des résumés PubMed et des articles en texte intégral. Des études ont démontré que BioGPT offre des performances de pointe dans diverses tâches NLP biomédicales, telles que l'extraction des entités nommées (NER), la génération de texte et la réponse aux questions biomédicales[39]. La figure 1.11 présente la structure de BioGPT lorsqu'il est adapté à des tâches ultérieures.

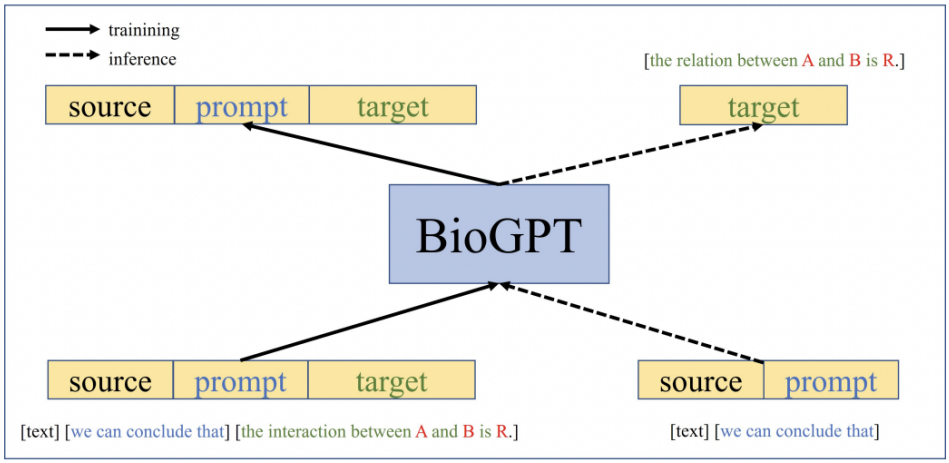


Figure 1.11: Structure de BioGPT lors de son adaptation à des tâches spécifiques [39].

L'entraînement du modèle BioGPT a été effectué à partir de zéro en utilisant un vaste ensemble de données comprenant 15 millions d'articles de recherche provenant de PubMed. Ce modèle a été évalué sur six tâches de traitement du langage naturel biomédical et a démontré des performances supérieures à celles de GPT-2 en ce qui concerne la génération de texte biomédical. De plus, BioGPT a surpassé les résultats de l'état de l'art sur trois tâches d'extraction de relations de bout en bout et une tâche de question-réponse[40][41].

Une étude de cas spécifique à la génération de texte biomédical a également démontré la capacité de BioGPT à produire des descriptions fluides pour les termes biomédicaux[42]. Ces résultats indiquent que BioGPT est particulièrement adapté à la génération de texte médical, surpassant ainsi les performances de GPT-2 dans ce domaine.

### 1.5.3 Notions sur les Ontologies

- **UMLS:** (Unified Medical Language System) est une ontologie contenant des informations exhaustives sur les termes médicaux. En reconnaissant les entités dans le texte qui correspondent à des termes médicaux, il est possible d'obtenir des informations complémentaires telles que la définition, les synonymes, les relations avec d'autres termes[10], la figure ci-dessous1.12 présente l'architecture de l'ontologie UMLS.

[9]https://github.com/microsoft/BioGPT
[10]https://sites.nlm.nih.gov/research/umls/sourcereleasedocs/index.html

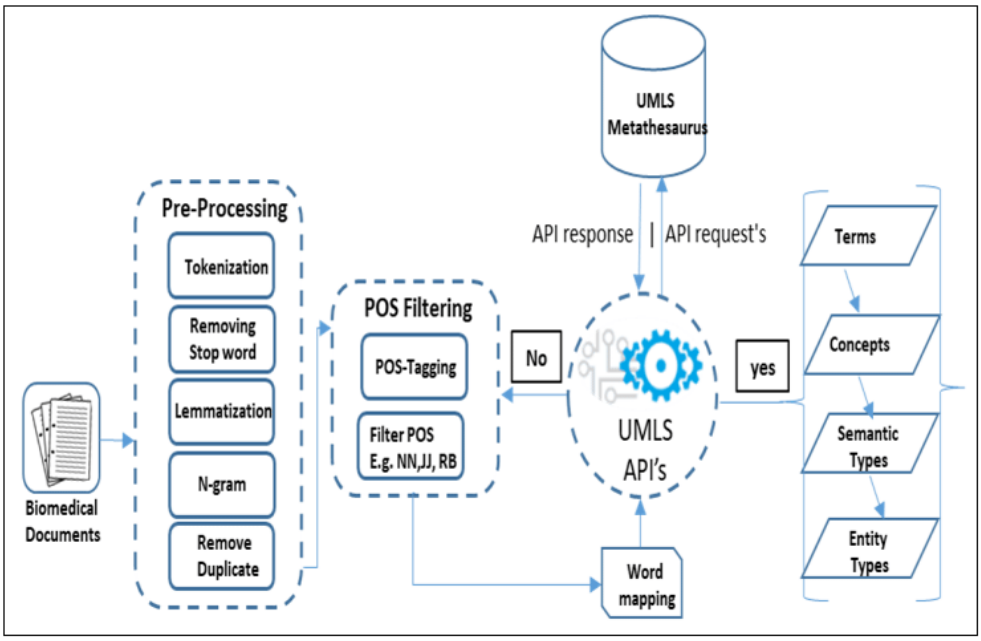


Figure 1.12: Représentation de l'architecture de l'ontologie UMLS[43].

- **SNOMED CT US:** (Systematized NOmenclature of MEDicine - Clinical Terms United States) est la source officielle de la SNOMED CT utilisée dans les systèmes de soins de santé aux États-Unis, et est une terminologie clinique qui regroupe la SNOMED Reference Terminology (SNOMED RT) et la Version 3 de la United Kingdom's Clinical Terms (anciennement connue sous le nom de Read Codes). Elle est organisée en une hiérarchie de 18 catégories de premier niveau. Ces entités, dites majeures, sont regroupées autour d'une racine appelée Top. Dans sa version de Juillet 2006, SNOMED CT totalise plus de 300 000 concepts et 770 000 descriptions en anglais[11].

  Le diagramme1.13 ci-dessus illustre la hiérarchie des concepts détaillés du terme médical "ulcer" dans SNOMED.

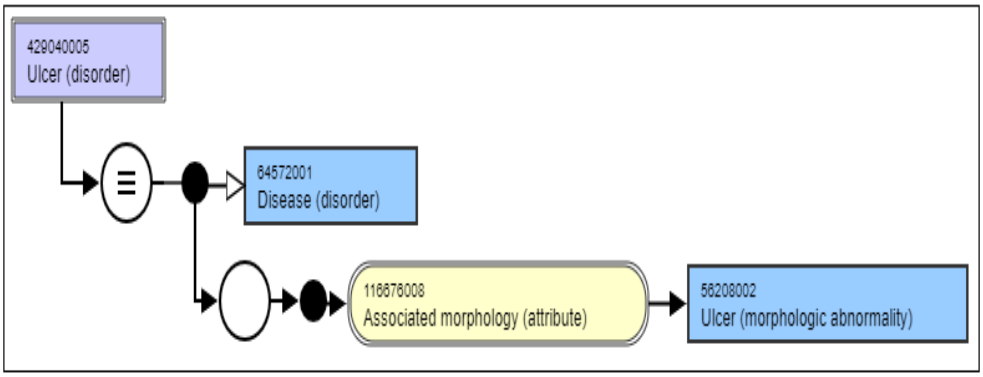


Figure 1.13: Hiérarchie des concepts détaillés du terme médical "ulcer".

[11]https://www.nlm.nih.gov/healthit/snomedct/us-edition.html

- **HPO:** (Human Phenotype Ontology) est un vocabulaire standardisé des anomalies phénotypiques dans les maladies humaines. Il est développé à partir de la littérature médicale, d'Orphanet, DECIPHER et OMIM. Le HPO contient plus de 13 000 termes et 156 000 annotations de maladies héréditaires. Il est utilisé pour les diagnostics différentiels basés sur le phénotype, les diagnostics génomiques et la recherche translationnelle. Le HPO fait partie de l'initiative Monarch, qui vise à intégrer les données biomédicales pour améliorer la recherche.[12], la figure1.14 suivante est un exemple d'une structure arborescente hiérarchique de données dans l'Human Phenotype Ontology (HPO):

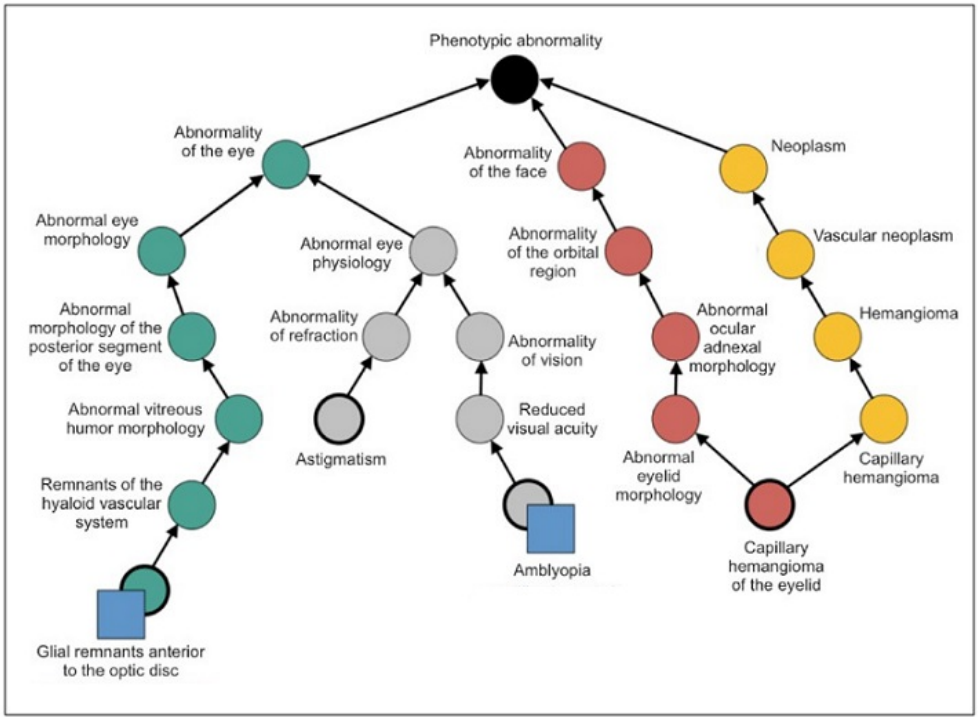


Figure 1.14: Exemple d'une structure arborescente hiérarchique de données dans HPO[44].

- **MeSH:** (Medical Subject Headings) est un thesaurus complet conçu par la NLM (National Library of Medicine), il est utilisé en particulier par PubMed (l'interface de la NLM à la base de données MEDLINE[13]). Il recense les concepts médicaux, les termes associés, les synonymes, les relations sémantiques et d'autres informations utiles pour l'analyse de données médicales. En détectant une entité dans le texte correspondant à un concept médical, MeSH peut fournir des informations complémentaires sur ce concept, notamment sa définition, ses caractéristiques ainsi que d'autres données importantes pour l'analyse médicale[14].

[12]https://hpo.jax.org/app/

[12]https://meshb.nlm.nih.gov./

[14]MEDLINE est une base de données bibliographiques qui couvre tous les domaines médicaux de l'année 1966 à nos jours: plus de 11 millions de références issues de 4 300 périodiques en langue anglaise.

Lorsque nous examinons le terme médical "ulcère" à travers le site officiel de MeSH, des informations spécifiques sur le terme ainsi que sa définition, représentée par le Scope note, sont fournies. L'action est illustrée dans l'image (figure 1.15) suivante:

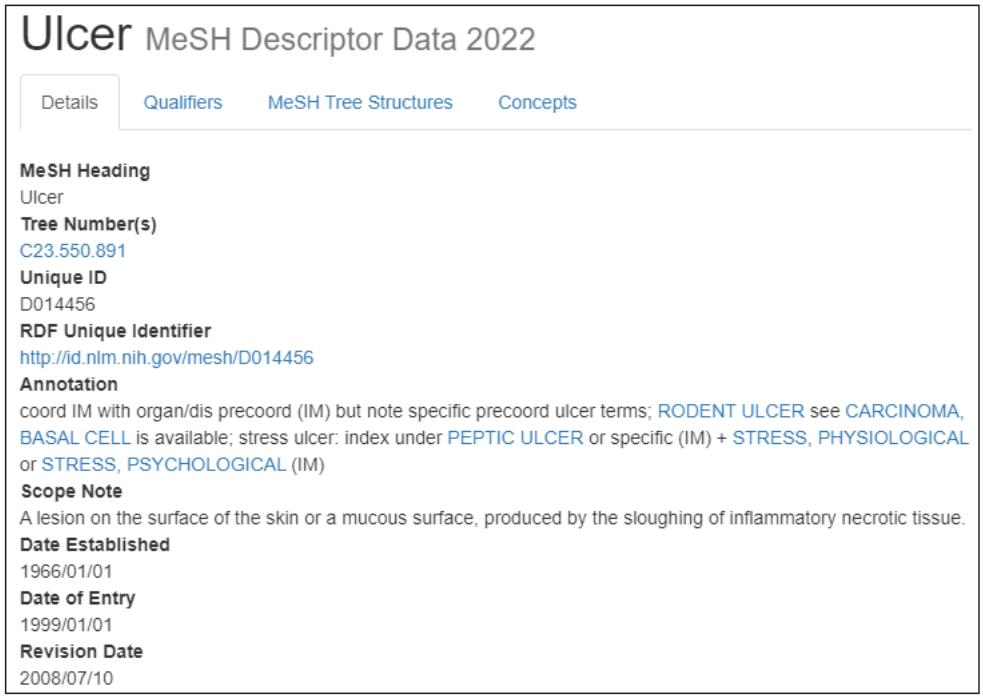


Figure 1.15: Illustration de l'extraction d'informations spécifiques du terme médical "ulcère" à partir du site officiel de MeSH.

- **WordNet:** est une base de données lexicale sémantique exhaustive qui contient des informations précises sur les mots de la langue anglaise, incluant leurs définitions, leurs synonymes, leurs antonymes, ainsi que d'autres propriétés lexicales et sémantiques. En identifiant une entité dans le texte qui correspond à un mot de la langue anglaise, WordNet peut fournir des informations complémentaires sur ce mot, telles que ses synonymes, ses antonymes, ses définitions, ainsi que d'autres relations sémantiques avec d'autres mots de la langue anglaise[15],la figure ci-dessous1.16 illustre l'architecture la base de données WordNet.

[15]https://wordnet.princeton.edu/

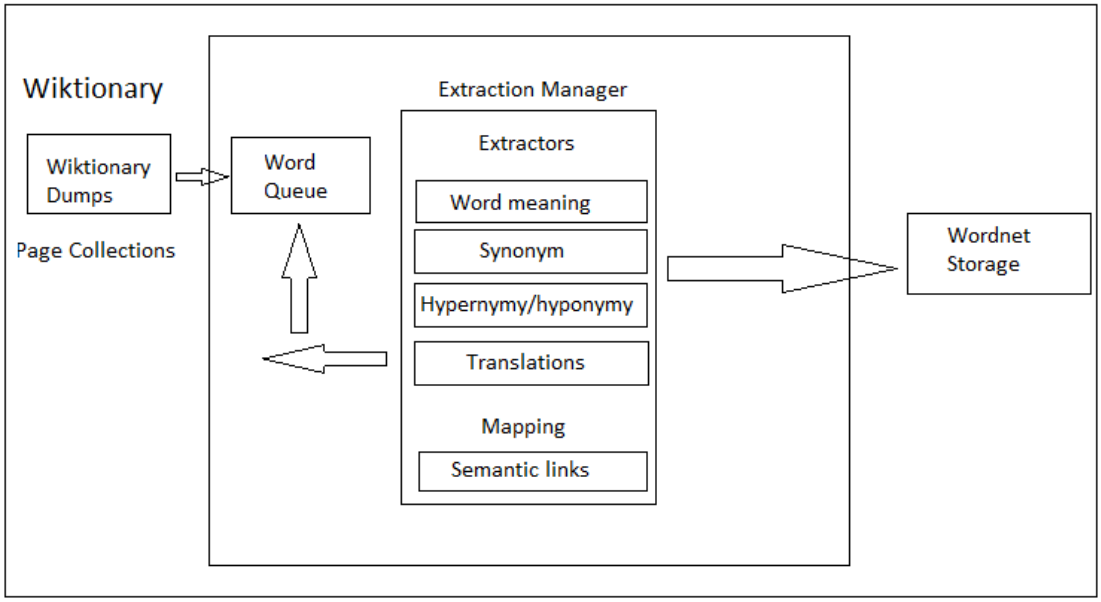


Figure 1.16: Architecture de la base de données lexico-sémantique WordNet[16].

Dans la section suivante, nous examinerons les travaux connexes concernant l'automatisation de la simplification des textes médicaux.

## 1.6 Travaux Connexes

La simplification des textes médicaux est devenue un enjeu central dans le domaine de la communication en santé, en raison de son rôle crucial dans l'amélioration de la compréhension des patients vis-à-vis des informations liées à leur santé. Au cours des dernières années, de nombreux travaux ont été réalisés pour simplifier les textes médicaux en utilisant diverses techniques, telles que l'approche sémantique, le résumé automatique, ainsi que la simplification lexicale et grammaticale. Ces approches novatrices ont montré des résultats encourageants quant à l'optimisation de la communication en santé et l'amélioration de la compréhension des patients.

Le premier effort vers la simplification automatisée est une enquête sur la simplification automatique des textes, une recherche est faite par Matthew Shardlow [45] qui modifie la syntaxe et le lexique du texte pour améliorer sa compréhensibilité pour les utilisateurs finaux. Elle couvre diverses approches de la simplification de texte, notamment les techniques lexicales, syntaxiques, de traduction automatique statistique et hybrides. L'enquête explore également les applications de la simplification de texte, telles que l'assistance aux apprenants de langue seconde et l'amélioration des technologies d'assistance. De plus, elle aborde les défis actuels auxquels est confronté le domaine de la simplification de texte.

[16]https://upload.wikimedia.org/wikipedia/commons/0/0e/Wiktionary-wordnet-creation-architecture.png

Dans le processus de simplification de texte, plusieurs modifications peuvent être apportées simultanément, tel qu'illustré par la figure1.17. En effet, dans l'exemple considéré, le terme perched est remplacé par sat, tandis que le terme roosted est éliminé lors d'une phase de désambiguïsation sémantique, étant donné qu'il ne correspond pas au contexte lexical du mot cat.

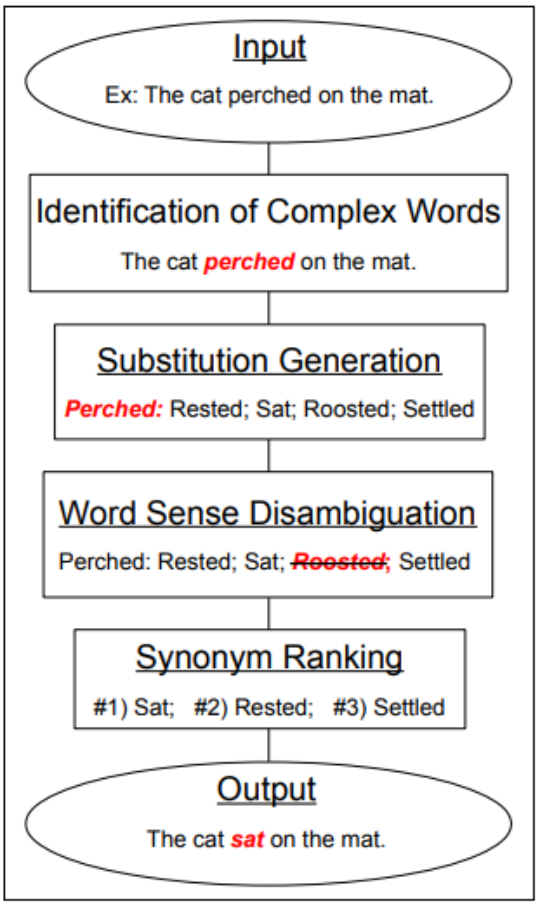


Figure 1.17: Processus de simplification lexicale[24].

Une autre recherche scientifique explore pour parvenir à des résultats concrets de l'utilisation de modèles de traduction automatique neuronale (TAN) afin d'améliorer la lisibilité des documents d'éducation destinés aux patients souffrant d'une faible littératie en santé. David Oniani et al.[46] constatent que de nombreux patients éprouvent des difficultés à comprendre les informations relatives à leur santé en raison de la complexité et du jargon du vocabulaire médical, ce qui peut avoir des conséquences néfastes sur leur santé. Ils proposent donc l'utilisation de modèles de TAN pour traduire le langage médical complexe en langage simple et accessible dans les documents d'éducation des patients, ce qui faciliterait leur compréhension pour les personnes ayant une faible littératie en santé. Par ailleurs, en élaborant un ensemble de données annotées nommé Silver Standard pour l'entraînement des modèles de TAN, puisqu'il n'existe pas d'ensemble de données public disponible pour cette tâche, ces auteurs examinent les méthodes d'évaluation de la difficulté

grammaticale et les méthodes de simplification grammaticale pour améliorer la lisibilité des informations de santé.

L'architecture du Transformer basé sur le modèle TAN est présentée dans la figure1.18 ci-dessous.

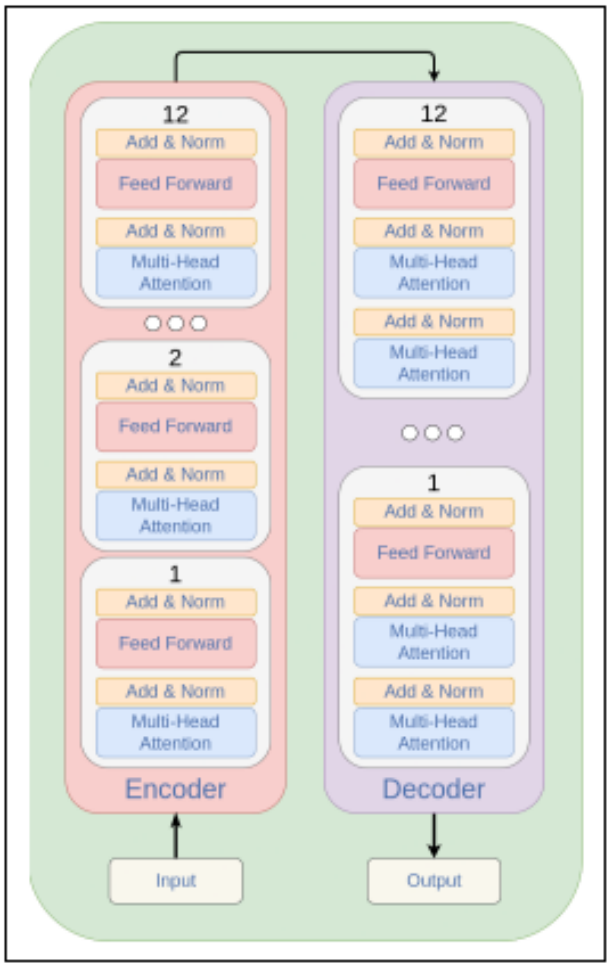


Figure 1.18: Transformer basé sur le modèle TAN[25].

Zhang et al. proposent dans un autre travail[47], une méthode innovante pour simplifier les phrases et les rendre plus accessibles. Ils combinent un modèle encodeur-décodeur avec un cadre d'apprentissage par renforcement profond pour optimiser le processus de simplification tout en préservant la signification originale, en utilisant divers ensembles de données, y compris Newsela, en nommant cette approche DRESS (Deep Reinforcement Sentence Simplification). En outre, les auteurs discutent des défis de la simplification des phrases et des limites des approches existantes et affirment que leur méthode était plus efficace car elle prend en compte à la fois les caractéristiques linguistiques et sémantiques des phrases, et utilise une fonction de récompense pour encourager des sorties simples, fluides et qui préserve la signification de l'entrée.

Ce modèle de simplification par renforcement profond comme l'illuste la figure1.19

suivante, est représenté par X pour la phrase complexe, Y pour la phrase de référence (simple) et Y^ pour la séquence d'actions (simplification) produite par le modèle encodeur-décodeur.

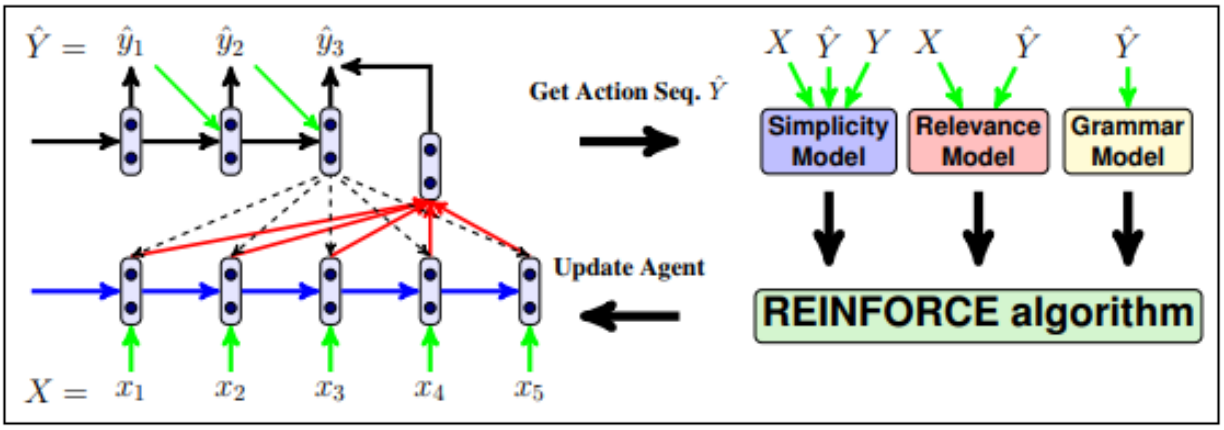


Figure 1.19: DRESS (Deep reinforcement learning simplification model)[26].

Une étude récente expose un modèle de simplification de phrases basé sur l'arbre appelé le modèle de simplification de phrases (MST), qui englobe la division, la suppression, la réorganisation et la substitution de mots et de phrases[48]. Pour entraîner leur modèle, Gurevych et al. rassemblent un grand ensemble de données complexes-simples appelé PWKP à partir de la version anglaise simple et de la version anglaise de Wikipédia. L'entraînement itératif du modèle se réalise en utilisant un algorithme appelé algorithme d'espérance-maximisation (AEM), donc ils proposent une méthode de cartographie de mots monolingues pour accélérer considérablement le processus d'entraînement. Cependant, les performances de leur modèle sont évaluées par rapport à plusieurs systèmes de base en utilisant diverses mesures telles que les scores de lisibilité et de fluidité. Les résultats révélent que leur modèle surpassait les systèmes de base en termes de scores de lisibilité.

La figure1.20 présente deux arbres: l'arbre d'entraînement (à gauche) et l'arbre de décodage (à droite). Ces représentations visuelles permettent de mieux comprendre le fonctionnement du modèle et son processus d'apprentissage.

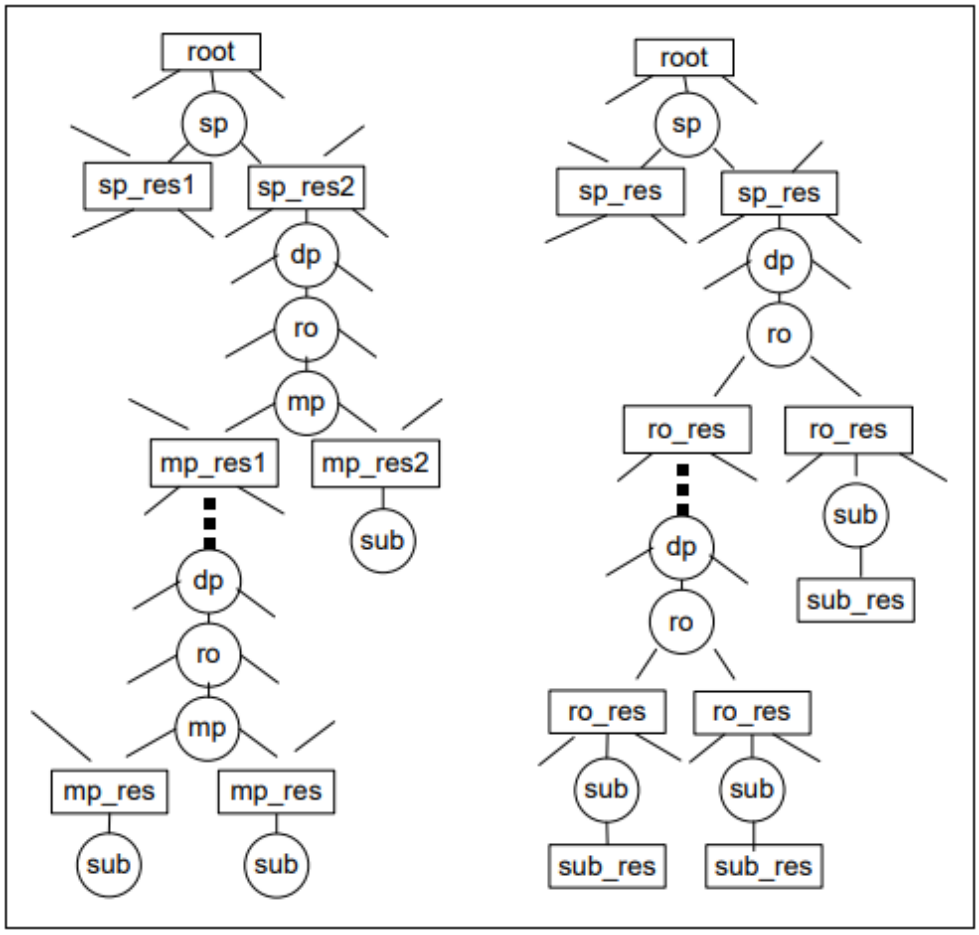


Figure 1.20: Arbre d'entraînement (à gauche) et l'arbre de décodage (à droite)[27].

Dans une autre étude similaire réalisée dans le même contexte, aborde la simplification automatique des textes médicaux dans le but de les rendre plus accessibles. L'auteur Remi Cardon[49] réalise une étude en utilisant trois ensembles de données médicaux comparables en langue française, le premier est le Cochrane dataset[17]: un corpus technique contenant des revues systématiques d'interventions de soins de santé, le deuxième est le corpus Medicaments[18]: un corpus simplifié contenant des informations sur les médicaments à l'intention des patients, et le dernier est le corpus Encyclopédie[19]: un corpus simplifié contenant des informations médicales générales à l'intention des patients.

[6]https://www.cochranelibrary.com
[7]https://base-donnees publique.medicaments.gouv.fr/
[19]https://fr.vikidia.org/

De plus de ces ensembles de données, l'auteur utilise deux ressources: une terminologie spécialisée, Snomed International[20], et un lexique généraliste issu du Wiktionary[21]. L'objectif était d'analyser et de modifier les textes médicaux en appliquant des techniques de simplification lexicale. Pour cela il effectue un alignement manuel des phrases des corpus comparables, ce qui a permis d'obtenir des données de référence pour analyser les processus de simplification utilisés.

L'article présente plusieurs contributions majeures, notamment la création de corpus comparables comprenant à la fois des textes médicaux techniques et simplifiés, l'observation des processus de simplification présents dans ces corpus, ainsi que l'évaluation des résultats selon trois critères de jugement: la grammaticalité, la simplification et la préservation sémantique. Les figures1.21 et 1.22 illustrent différents exemples de substitutions réalisées à l'aide des ressources disponibles. La figure1.21 présente des substitutions réussies où la sémantique des phrases reste fidèle aux phrases d'origine grâce à l'utilisation de synonymes tels que (absorption ; ingestion), (traitement ; prescription) ou (traiter ; soigner). En revanche, la figure1.22 présente des substitutions non réussies où la sémantique des phrases n'est pas préservée. Par exemple, il y a un changement sémantique dans le cas des synonymes (corps ; mort). En ce qui concerne les exemples avec les synonymes (main ; pince), (dents ; chicots) ou (tête ; citron), il s'agit de synonymes appartenant à différents niveaux de langage (normé ; jargon). Bien que cela n'altère pas considérablement la sémantique des phrases, la formulation devient plus familière, ce qui n'était pas l'effet recherché.

| *Avant substitution* | *Après substitution* |
|---|---|
| La nourriture n'a pas d'effet sur l'absorption d'anastrozole. | La nourriture n'a pas d'effet sur l'ingestion d'anastrozole. |
| Vous devez discuter avec votre médecin sur les risques et les options de traitement. | Vous devez discuter avec votre médecin sur les risques et les options de prescription. |
| Votre médecin peut vous prescrire un médicament visant à prévenir ou traiter cette perte osseuse. | Votre médecin peut vous prescrire un médicament visant à prévenir ou soigner cette perte osseuse. |

Figure 1.21: Exemple de substitutions réussies[28].

| *Avant substitution* | *Après substitution* |
|---|---|
| Un abcès est une accumulation de pus sous la peau ou à l'intérieur du corps. | Un abcès est une accumulation de pus sous la peau ou à l'intérieur du mort. |
| Syndrome du canal carpien (fourmillement, douleur, sensation de froid, faiblesse dans certaines parties de la main). | Syndrome du canal carpien (fourmillement, douleur, sensation de froid, faiblesse dans certaines parties de la pince). |

Figure 1.22: Exemple de substitutions non réussies[28].

[9]https://www.snomed.org/
[21]https://fr.wiktionary.org/

## 1.7 Discussion

Notre étude se consacre à l'exploration des divers travaux relatifs à l'automatisation de la simplification des textes dans le domaine médical. Le tableau ci-dessous1.3 récapitule l'ensemble des travaux mentionnés précédemment.

| **Travaux** | **Problème résolu** | **Méthodes utilisées** | **Corpus utilisés** |
|---|---|---|---|
| Shardlow et al.[45] | Faciliter la compréhension du langage médical complexe pour les utilisateurs finaux. | Techniques basées sur des règles et d'apprentissage profond | Corpus personnel |
| Oniani et al. [46] | La faible littératie en santé dans les supports d'éducation des patients. | Neural machine translation (NMT) models | Silver standard |
| Zhang et al. [47] | Simplification des textes médicaux pour améliorer la lisibilité et la compréhension. | DRESS Neural Machine translation | Newsela[22] |
| Gurevych et al. [48] | Transformation des phrases complexes en phrases plus simples tout en préservant leur sens. | Tree-based Simplification Model (TSM), | Wikipedia PWKP[23] |
| Remi Cardon. [49] | Simplification automatique des textes médicaux pour améliorer leur lisibilité et leur accessibilité. | Substitution lexicale à l'aide de ressources existantes et l'alignement manuel des phrases. | Cochrane[24] et Encyclopedie [25] |

Table 1.3: Etudes menées concernant la simplification des textes médicaux.

[11]https://newsela.com
[12]http://simple.wikipedia.org
[13]https://www.cochranelibrary.com
[25]https://fr.vikidia.org/

## 1.8 Conclusion

Dans ce chapitre, nous avons introduit la simplification des textes médicaux et comment cette tâche est devenue de plus en plus cruciale dans le domaine de la médecine. Nous avons également discuté des différentes approches existantes mentionnées dans certaines études réalisées pour simplifier les textes médicaux, ainsi que des défis liés à cette tâche. Dans les chapitres suivants, nous décrirons en détail notre contribution à la simplification automatique des textes médicaux.

# Chapitre 2

# Conception

## 2.1 Introduction

*"La simplification des textes médicaux est un enjeu majeur pour l'amélioration de la qualité des soins et de la sécurité des patients. Elle permet d'optimiser la compréhension des informations médicales par les patients et les professionnels de santé."*, Jean-Luc Dubois[50].

Dans ce chapitre, Nous aborderons la conception de notre projet avec l'approche sémantique pour la simplification des textes médicaux. Cette approche repose sur une combinaison de techniques de traitement du langage naturel et d'ontologies médicales afin d'identifier les concepts complexes et leurs relations dans le texte, pour ensuite les remplacer par des termes plus simples et couramment utilisés. La méthode comprend plusieurs étapes, notamment le prétraitement du texte, l'analyse sémantique, la simplification des termes, le résumé et la génération de texte.

## 2.2 Schéma Global de la Solution Proposée

La figure2.1 suivante offre une vue d'ensemble complète de notre projet basée sur une approche sémantique globale qui comprend trois solutions distinctes. La première solution consiste en la collecte et le prétraitement des données, ainsi que la reconnaissance des données et la génération de texte. La deuxième solution se focalise principalement sur le résumé de texte. Ces deux solutions initiales sont ensuite complétées par la troisième solution, qui se concentre sur la simplification des textes. Cette dernière solution intègre également des étapes d'analyse sémantique, d'évaluation et de déploiement du modèle, et peut être mise en œuvre de manière indépendante des deux premières solutions. Ainsi, bien que les deux premières solutions soient indispensables pour fournir les données de base, la troisième solution offre une approche distincte pour simplifier les textes en utilisant les mêmes étapes d'analyse sémantique, d'évaluation et de déploiement du modèle.

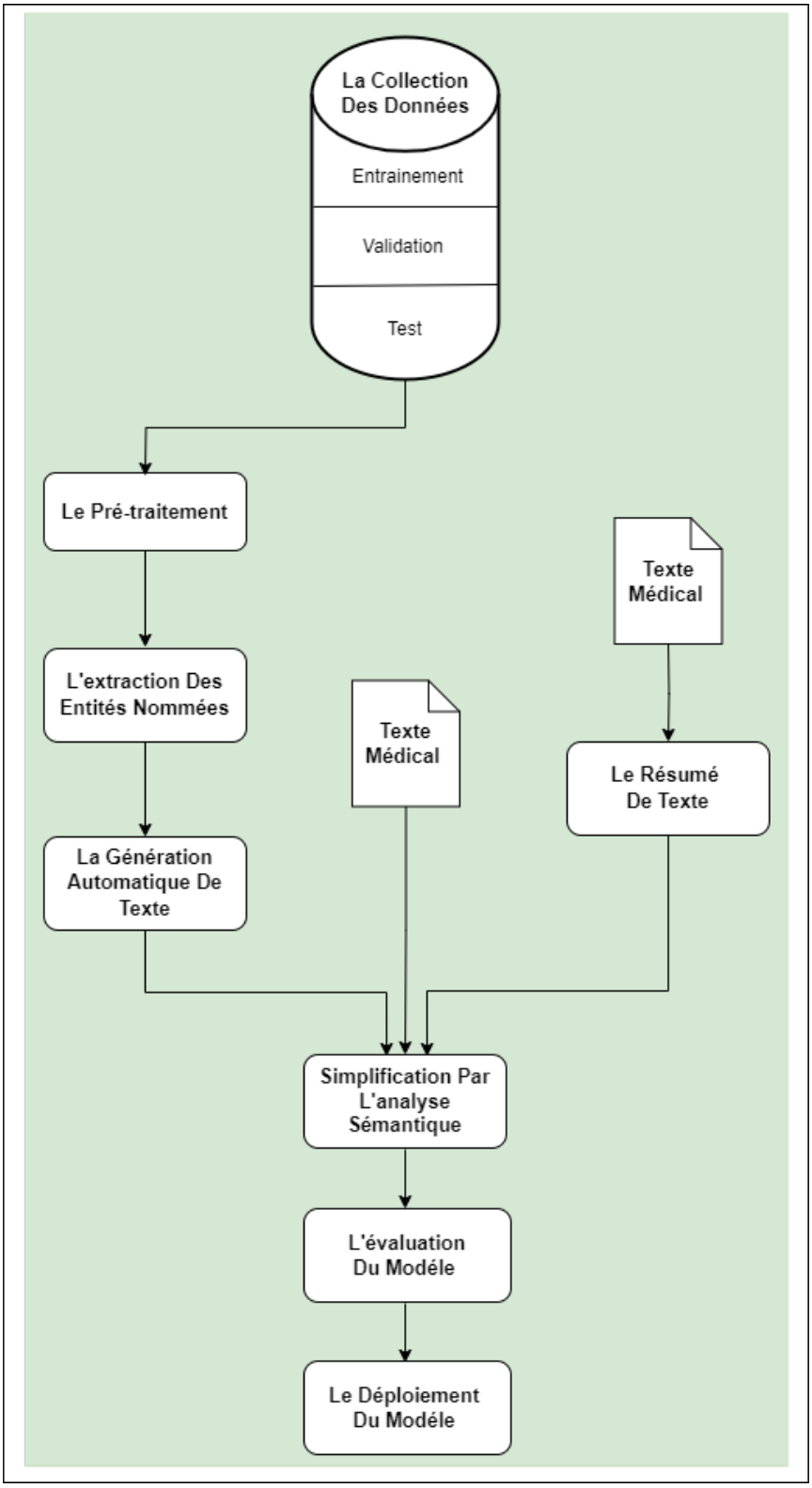


Figure 2.1: Aperçu du schéma global basé sur l'approche sémantique.

Par la suite, nous détaillerons chaque étape séparément.

## 2.3 Collection des Données

Dans le cadre de ce projet, nous utilisons l'ensemble de données Cochrane, qui est accessible via le site Hugging Face[1]. Cet ensemble de données utilisé pour simplifier les paragraphes de textes médicaux s'agit d'une collection qui recense des revues systématiques de questions cliniques comportant de nombreux résumés rédigés en anglais simple. Cette collection de données Cochrane comprend environ 4 500 paires de phrases, qui sont converties du format JSON au format TXT pour faciliter leur exploitation. Les données sont réparties en trois catégories, à savoir: pour l'entraînement 3568 fichiers, pour la validation 411 fichiers et 480 fichiers pour le test.

Conformément aux informations fournies par GEM sur son site officiel[2], le jeu de données Cochrane a été ajouté par Ashwin Devaraj de l'Université du Texas à Austin. Ce projet a été financé grâce à une subvention du National Institutes of Health (NIH)[51]. L'objectif principal de ce jeu de données est de développer un modèle capable de simplifier les textes médicaux afin de les rendre plus accessibles aux lecteurs non spécialisés dans le domaine médical. Aucune annotation supplémentaire n'a été réalisée pour ce jeu de données. Les champs de données comprennent l'identifiant unique de l'exemple (gem-id), l'identifiant DOI de la revue Cochrane à partir de laquelle l'exemple a été généré, la source (un extrait d'une revue Cochrane) et la cible (un résumé en langage clair d'une revue Cochrane correspondant approximativement au texte source, tout en étant plus court en termes de nombre de mots). Les deux textes semblent revêtir un caractère scientifique dans leur style rédactionnel, ce qui peut rendre leur compréhension difficile pour les lecteurs non familiarisés avec le domaine médical.

La figure2.2 suivante est un exemple de source extrait de l'ensemble de données Cochrane.

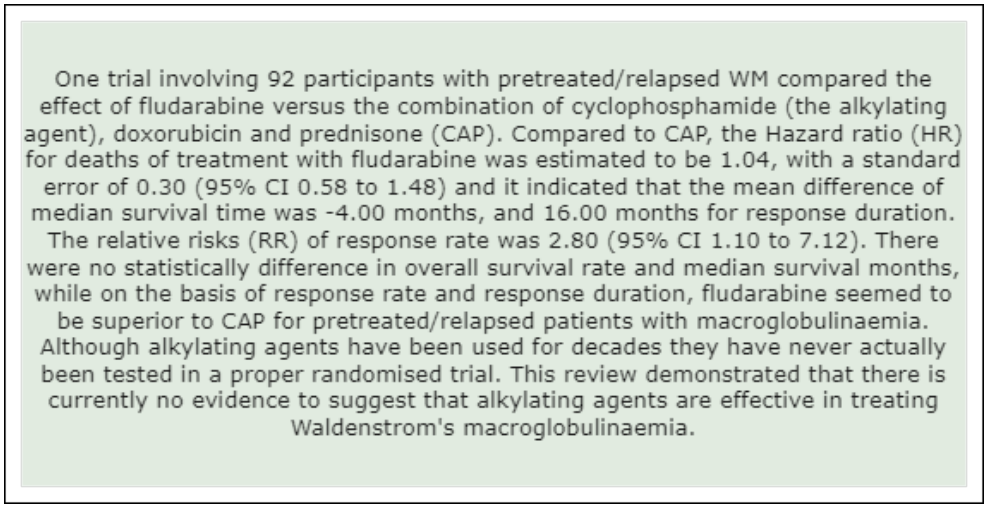


Figure 2.2: Exemple d'un texte source extrait de l'ensemble de données Cochrane.

Par contre, la figure2.3 ci-dessous est un exemple de cible extrait de l'ensemble de

[1]https://huggingface.co/datasets/GEM/cochrane-simplification
[2]https://gem-benchmark.com/data-cards/cochrane-simplification

données Cochrane.

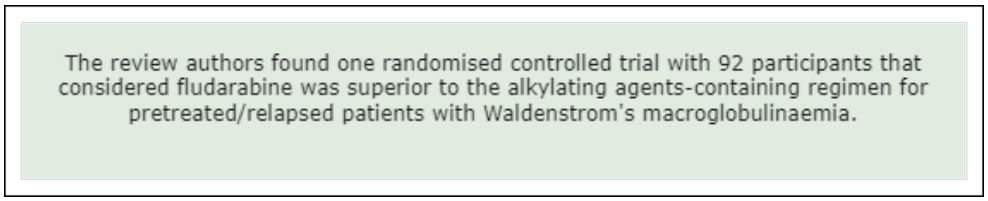


Figure 2.3: Exemple d'un texte cible extrait de l'ensemble de données Cochrane.

Bien que ce constat ne provienne pas du site officiel de l'ensemble de données, nous pouvons observer grâce au nombre élevé de téléchargements, atteignant 419 téléchargements[3], que ce jeu de données Cochrane est largement utilisé par les développeurs intéressés par le domaine médical.

Dans le cadre de notre étude, les deux sections, à savoir la source et la cible, seront extraites afin de les utiliser ultérieurement dans nos travaux. La figure2.4 suivante présente un exemple de texte médical de l'ensemble de données Cochrane en anglais.

Two trials met the inclusion criteria. One compared 2% ketanserin ointment in polyethylene glycol with PEG alone, used twice a day by 40 participants with arterial leg ulcers, for eight weeks or until healing, whichever was sooner. One compared topical application of blood-derived concentrated growth factor with standard dressing (polyurethane film or foam); both applied weekly for six weeks by 61 participants with non-healing ulcers (venous, diabetic arterial, neuropathic, traumatic, or vasculitic). Both trials were small, reported results inadequately, and were of low methodological quality. Short follow-up times (six and eight weeks) meant it would be difficult to capture sufficient healing events to allow us to make comparisons between treatments. One trial demonstrated accelerated wound healing in the ketanserin group compared with the control group. In the trial that compared CGF with standard dressings, the number of participants with diabetic arterial ulcers were only reported in the CGF group (9/31), and the number of participants with diabetic arterial ulcers and their data were not reported separately for the standard dressing group. In the CGF group, 66.6% (6/9) of diabetic arterial ulcers showed more than a 50% decrease in ulcer size compared to 6.7% (2/30) of non-healing ulcers treated with standard dressing.

Figure 2.4: Exemple de texte médical de l'ensemble de données Cochrane[4].

Au sein de cet extrait provenant de la section d'entraînement de l'ensemble de données Cochrane, on observe l'utilisation de plusieurs termes médicaux et techniques afin de décrire les essais cliniques. Parmi les termes médicaux, nous pouvons identifier des expressions telles que "ulcères artériels", "ulcères veineux, artériels diabétiques, neuropathiques, traumatiques ou vasculitiques", "guérison accélérée des plaies" et "diminution de plus de 50

[3]En mois de Mai 2023.

[4]https://huggingface.co/datasets/GEM/cochrane-simplification

pour cent de la taille de l'ulcère". Quant aux termes techniques, il convient de mentionner l'utilisation de formulations telles que "2 pour cent de pommade de kétransérine dans du polyéthylène glycol", "application topique de facteur de croissance concentré dérivé du sang", "pansement standard (film ou mousse de polyuréthane)", "événements de guérison suffisants pour permettre des comparaisons entre les traitements", "suivi à court terme de six et huit semaines", "nombre de participants", "qualité méthodologique faible" et "résultats rapportés de manière insuffisante".

Tous ces termes, qui peuvent être difficilement compris par des non-spécialistes du domaine, servent à caractériser les conditions médicales étudiées, à appréhender les résultats obtenus ainsi que les limitations de l'étude.

## 2.4 Pré-traitement des Données

Le prétraitement est le processus de nettoyage et de transformation des données brutes avant qu'elles ne soient utilisées pour l'analyse ou l'entraînement des modèles.

Dans notre projet, nous avons identifié plusieurs étapes de prétraitement des données en fonction de nos objectifs spécifiques. Nous avons retenu les étapes suivantes, telles qu'illustrées dans la figure2.5 ci-dessous, car elles sont cruciales pour préparer les données de manière adéquate avant de les utiliser, ce qui permet d'optimiser les performances et la qualité des résultats obtenus.

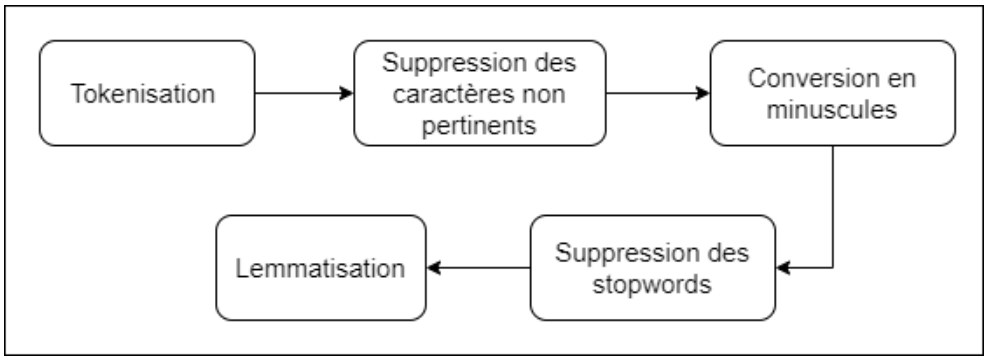


Figure 2.5: Aperçu des étapes utilisées dans la phase de pré-traitement des données.

Par la suite, nous procéderons à une description détaillée de chaque étape.

### 2.4.1 Tokenisation

Pendant cette étape, le texte est subdivisé en unités plus petites appelées "token" ou "jeton" en français. Un jeton fait référence à un élément linguistique[52], tel qu'un mot ou une phrase, pouvant être analysé de manière indépendante. Cette notion est illustrée dans la figure2.6 présentée ci-dessus.

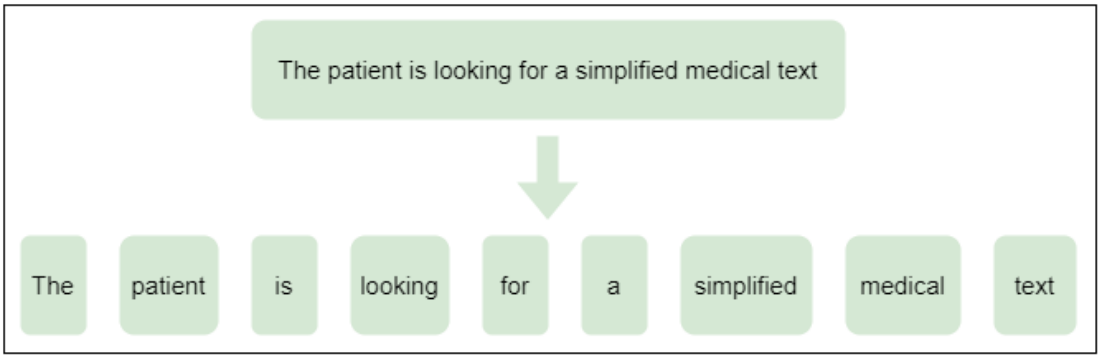


Figure 2.6: Représentation de l'étape de tokenisation.

### 2.4.2 Nettoyage des Données

L'étape de nettoyage des données consiste à identifier et à corriger les erreurs et les incohérences dans un jeu de données, afin d'en améliorer la qualité et la fiabilité pour une analyse ultérieure. Cette étape elle-même comprend plusieurs sous-étapes, on commence par:

**La Suppression des Caractères non Pertinents:** Dans cette étape, nous procédons au nettoyage de notre ensemble de données en supprimant les balises HTML, la ponctuation et les caractères non pertinents comme : @ + ! ?

**La Conversion en Minuscules:** C'est une technique de prétraitement des données textuelles qui consiste à convertir toutes les lettres d'un mot en minuscules, tel qu'illustré dans le schéma2.7 ci-dessous:

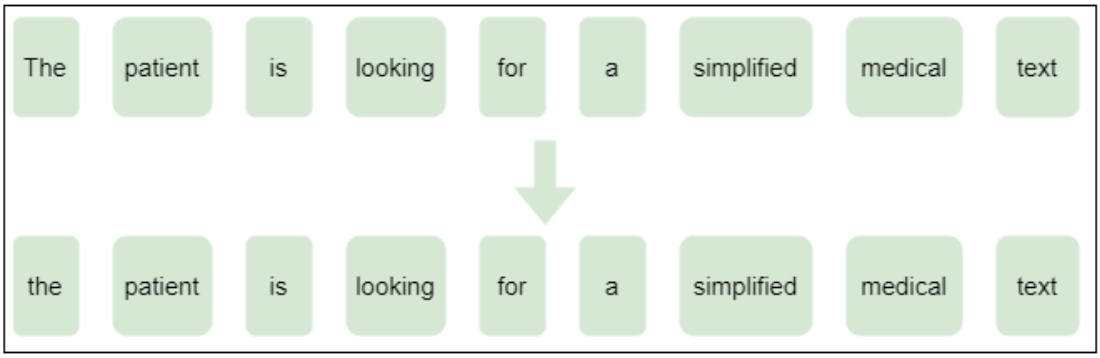


Figure 2.7: Représentation de l'étape de la conversion en minuscules.

**La Suppression des Stopwords:** Elle consiste à éliminer les mots courants et non informatifs qui n'apportent pas de valeur sémantique significative à l'analyse. Cette étape permet de réduire la dimensionnalité du texte et de se concentrer sur les mots clés et les informations pertinentes pour l'analyse ultérieure, comme indiqué dans l'illustration2.8 suivante:

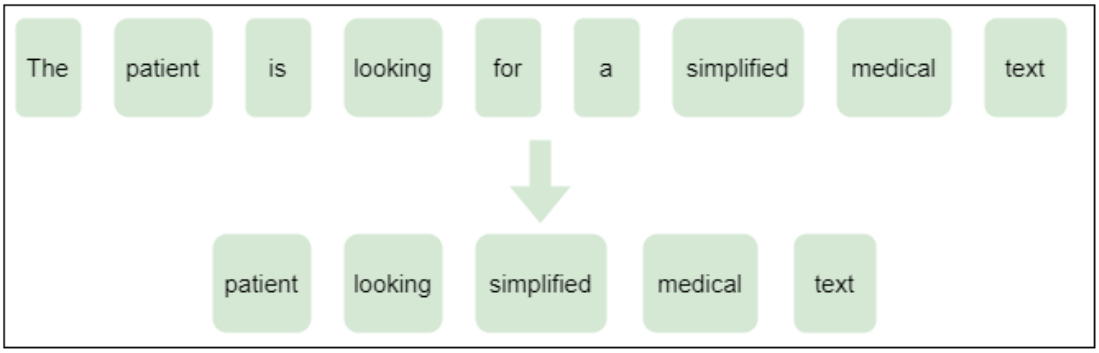


Figure 2.8: Représentation de l'étape de la suppression des stopwords.

**La Lemmatisation:** Elle consiste à réduire les mots à leur forme de base ou de racine. Par exemple, en utilisant la lemmatisation, le mot "jouer" serait ramené à son lemme "jouer", tandis que le mot "joue" serait réduit à "jouer" également, comme le montre la figue2.9 ci-après:

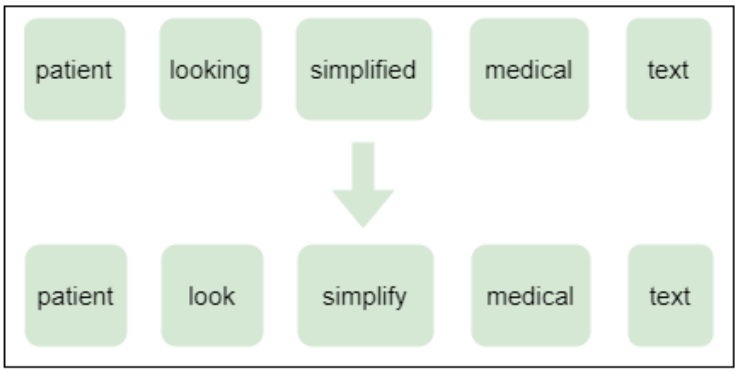


Figure 2.9: Représentation de l'étape de la lemmatisation.

Arrivées à cette étape, nous disposons d'un ensemble de données nettoyées, ce qui nous permet d'appliquer les traitements nécessaires.

## 2.5 Extraction des Entités Nommées

Pour une meilleure compréhension du contenu nous essayons d'extraire des entités telles que CANCER, CELL, ORGAN, ORGANISM, SIMPLE-CHEMICAL, TISSUE, etc, du modéle en-ner-bionlp13cg-md [5].

Dans la figure2.10 suivante, nous illustrons un texte après l'étape d'extraction des entités nommées médicales.

[5]https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.5.1/en-ner-bionlp13cg-md-0.5.1.tar.gz

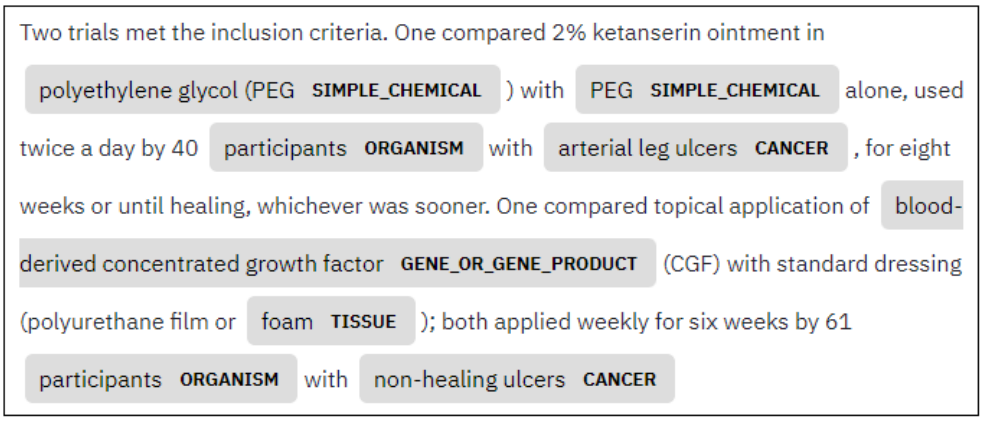


Figure 2.10: Exemple de texte après l'extraction des entités nommées médicales.

Après avoir passé le texte dans le modèle, nous remarquons que toutes les entités extraites sont mises en évidence, et devant chaque entité, son type spécifique est identifié. Cela nous permet d'obtenir une vue d'ensemble claire des différentes entités présentes dans le texte et de comprendre rapidement les types d'entités impliquées, tels que CANCER, GENE-OR-GENE-PRODUCT, ORGANISM, SIMPLE-CHEMICAL, TISSUE, etc.

## 2.6 Traitement Initial

Le traitement de texte englobe l'ensemble des techniques et méthodes employées pour la manipulation et l'analyse automatique du texte, généralement à l'aide de logiciels ou d'algorithmes informatiques. Il englobe un large éventail de tâches comprenant la reconnaissance de caractères, la correction orthographique et grammaticale, la segmentation de phrases, l'extraction d'informations, la traduction automatique, la génération de texte, la recherche d'informations, le résumé automatique, la classification de texte, ainsi que d'autres domaines connexes[53].

Ce travail présente trois approches: la première propose de générer du texte médical à partir de l'entrée de l'utilisateur, puis de le simplifier, tandis que la deuxième propose de résumer le texte d'entrée de l'utilisateur avant de le simplifier et la troisième solution consiste à simplifer le texte d'entrée de l'utilisateur.

### 2.6.1 Génération de Texte

Afin de créer automatiquement un texte cohérent et compréhensible par les machines, nous nous sommes intéressées à utliser les modèles de langage pré-entraînés. Il convient de noter que BioGPT a été entraîné spécifiquement sur un corpus de données biomédicales, ce qui lui confère un avantage par rapport à BioBERT qui a été pré-entraîné sur un ensemble de données biomédicales plus large et peut donc ne pas être aussi performant pour la génération de texte biomédical spécifique. Dans ce cas nous avons porté une attention

particulière à une architecture qui s'est révélée prometteuse: BioGPT. Ce modèle, basé sur les Transformers,est suscité un grand intérêt dans le domaine du traitement du langage naturel appliqué à la biologie et à la biomédecine.

Les chercheurs et développeurs ont la possibilité d'améliorer les performances de BioGPT dans des tâches spécifiques en l'affinant avec leurs propres ensembles de données. L'affinage ou le fine-tuning de BioGPT implique l'utilisation de sa version pré-entraînée et son adaptation à une tâche ou un domaine spécifique en le formant sur un ensemble de données plus restreint et spécifique à cette tâche.

Le tableau2.1 suivant représente un exemple d'un court texte médical et de sa version étendue générée à l'aide de BioGPT.

| **Texte d'entrée court** | **Version longue générée par BioGPT** |
|---|---|
| Le patient présentait une tachycardie et une dyspnée. Le traitement a consisté à administrer de l'épinéphrine et ... | Le patient présentait une tachycardie et une dyspnée. Le traitement a consisté à administrer de l'épinéphrine et du gluconate de calcium par voie intraveineuse, ce qui a permis une amélioration rapide à la fois des symptômes cliniques (pression artérielle) et des anomalies électrocardiographiques. |

Table 2.1: Exemple illustrant un texte médical avant et après la génération par BioGPT.

### 2.6.2 Résumé de Texte

La création d'un résumé de texte consiste à extraire les points importants d'un texte source et fournir une version condensée du texte qui conserve les informations clés.

Dans le cadre de notre travail, nous avons opté pour une approche extractive afin de générer un résumé sans créer de nouvelles phrases.

Différentes techniques sont utilisées pour réaliser le résumé, et après avoir effectué plusieurs tests, il semble que TextRank[54] soit la meilleure technique, par concéquent, nous avons choisi d'utiliser cette technique pour résumer nos textes en raison de sa simplicité et de son efficacité dans la génération de résumés extractifs. Avant d'approfondir notre approche, il est essentiel d'examiner en détail le flux de l'algorithme TextRank que nous avons suivi. Ce flux est présenté dans la figure2.11 suivante.

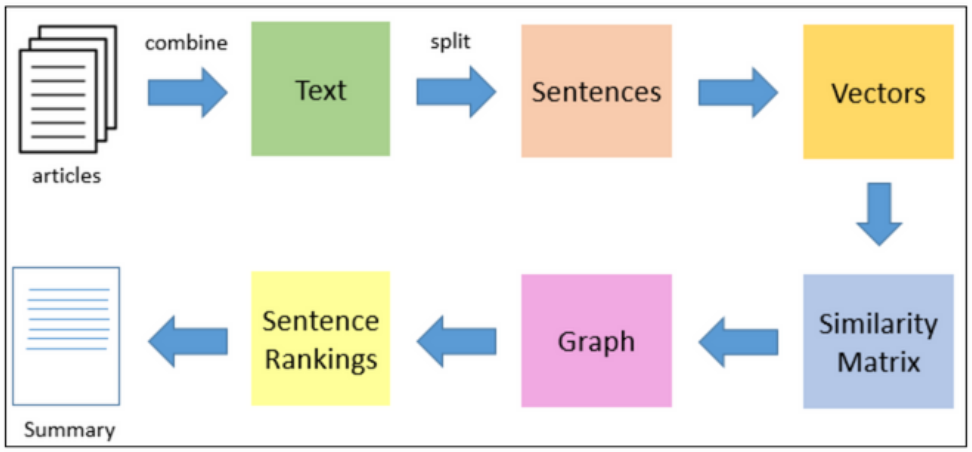


Figure 2.11: Architecture de l'algorithme TextRank[6].

Pour mettre en œuvre et mieux comprendre l'algorithme TextRank, il est essentiel de prendre en compte quelques points clés. Tout d'abord, le texte doit être prétraité en tokenisant les phrases tout en effectuant éventuellement une lemmatisation ou une normalisation, cette technique utilise ainsi des word embeddings, car ils permettent de représenter les mots sous forme de vecteurs numériques dans un espace continu. Les word embeddings capturent les similarités sémantiques. Ensuite, un graphe de co-occurrence est construit en représentant les mots comme des nœuds et les relations de co-occurrence entre les mots comme des arêtes. L'importance des mots est calculée à l'aide de l'algorithme PageRank, en prenant en compte la fréquence des co-occurrences et la position des mots dans le graphe. Enfin, les phrases les plus importantes, évaluées en fonction de l'importance de leurs mots constitutifs, sont sélectionnées pour former le résumé final.

Ci-dessous est présenté un tableau3.2 illustrant un exemple de texte médical long accompagné de sa version résumée.

| **Texte d'entrée long** | **Texte résumé** |
|---|---|
| La sclérose en plaques est une maladie auto-immune qui affecte le système nerveux central. Elle se caractérise par l'inflammation et la démyélinisation des nerfs, ce qui entraîne des problèmes de coordination, de mobilité et de sensibilité. Les symptômes de la sclérose en plaques peuvent varier d'une personne à l'autre et peuvent être traités avec des médicaments immunomodulateurs pour réduire les poussées et ralentir la progression de la maladie. | La sclérose en plaques est une maladie auto-immune du système nerveux central, marquée par l'inflammation et la démyélinisation des nerfs, entraînant des troubles de coordination, de mobilité et de sensibilité, mais qui peut être atténuée et ralentie grâce à l'utilisation de médicaments immunomodulateurs. |

Table 2.2: Exemple de texte avant et après avoir effectuer le résumé.

[6]https://cdn.analyticsvidhya.com/wp-content/uploads/2018/10/block3.png

### 2.6.3 Simplification de Texte

La simplification de texte consiste à rendre un texte plus compréhensible et accessible en le reformulant de manière plus simple, tout en préservant son sens principal. Cela implique généralement de réduire la complexité syntaxique, de remplacer des termes difficiles par des synonymes plus courants, de supprimer des informations non essentielles, etc. La simplification de texte est souvent utilisée pour rendre les écrits plus accessibles aux personnes ayant des difficultés de lecture, aux apprenants en langue seconde et à des publics spécifiques[55].

Le tableau2.3 ci-dessous présente un exemple de texte complexe ainsi que sa version simplifiée.

| **Texte avant la simplification** | **Texte après la simplification** |
|---|---|
| La consommation excessive de boissons sucrées est associée à un risque accru de développer des maladies chroniques telles que l'obésité, le diabète de type 2 et les maladies cardiovasculaires. | Boire trop de boissons sucrées augmente le risque de maladies chroniques comme l'obésité, le diabète de type 2 et les problèmes cardiaques. |

Table 2.3: Exemple de texte avant et après la simplification.

D'après plusieurs recherches, il a été constaté que ces trois domaines: la simplification de texte, la génération de texte ainsi que le résumé de texte se chevauchent fréquemment, renforçant l'idée que certains modèles peuvent être utilisés de manière polyvalente. Cependant, la présentation de ces résultats peut varier selon les études, par exemple, un modèle de génération de texte peut simplifier un texte en reformulant les phrases de manière plus accessible. De même, un modèle de résumé de texte peut englober la génération de texte pour créer un résumé concis. L'utilisation dépend donc du point de vue et des objectifs spécifiques de chaque tâche[50][56].

Cependant, malgré le traitement du texte, il reste certains mots complexes. Afin de remédier à cela, nous allons enrichir notre texte simplifié grâce à l'analyse sémantique en utilisant des ontologies et des ressources lexico-sémantiques.

## 2.7 Analyse Sémantique

L'analyse sémantique en médecine est un processus qui vise à comprendre et à interpréter le sens des textes médicaux en se basant sur le contexte clinique. Cela implique l'extraction des concepts médicaux, et l'interprétation des informations médicales pour en extraire des connaissances significatives[57].

L'analyse sémantique en médecine permet d'améliorer la compréhension des textes médicaux.

### 2.7.1 Enrichissement Sémantique

L'enrichissement sémantique se réfère à un processus visant à ajouter des informations sémantiques supplémentaires à un texte ou à des données dans le but d'améliorer leur compréhension, leur interprétation et leur utilisation. Les modalités de fonctionnement de l'enrichissement sémantique peuvent varier en fonction des techniques et des outils employés, cependant, cela implique généralement l'utilisation de ressources lexicales, de bases de connaissances ou de modèles sémantiques préexistants pour extraire des informations supplémentaires et les associer au contenu d'origine. L'enrichissement sémantique trouve de nombreuses applications, telles que l'amélioration de la recherche d'informations, l'assistance à la compréhension automatique des textes, l'interprétation des données, la recommandation de contenus, ainsi que la facilitation de l'intégration et de l'interopérabilité des systèmes d'information[58].

Les ontologies et les ressources lexico-sémantique tels que UMLS, MeSH, WordNet, etc, sont des ressources qui contiennent des informations structurées sur des concepts, des relations et des propriétés. Ces ressources peuvent être utilisées pour enrichir les entités identifiées dans le texte avec des informations supplémentaires, telles que des synonymes, des définitions, des relations sémantiques, etc[59].

Le pseudo-code de l'image2.12 suivante explique brièvement la fonction utilisée pour la simplification de text en utilisant les ontlogies.

```
Fonction termsDetails(mots_médicaux)
DEBUT
    dictionnaire = dictionnaire vide
    pour chaque terme dans mots_médicaux faire
        cui = conceptID(terme)
        si cui non null alors
            type_sémantique = semanticTypes(cui)
            définition = definition(cui)
            dictionnaire.ajouter({"Terme": terme, "ID de concept": cui,
            "Types sémantiques": type_sémantique, "Définition": définition})
        fin si
    fin pour
    retourner dictionnaire
Fin Fonction
```

Figure 2.12: Illustration de la simplification de text à travers un pseudo-code.

La figure précédente illustre le processus d'extraction des termes médicaux à partir de notre texte original, suivi de la phase de simplification à l'aide des ontologies. Notre fonction parcourt méticuleusement chaque terme de la liste des termes médicaux. Dans un premier temps, nous recherchons l'identificateur unique de concept (CUI) associé à chaque terme. Une fois le CUI identifié, nous recueillons le type sémantique du terme ainsi que sa définition. Ensuite, nous procédons à la construction d'un dictionnaire contenant le terme lui-même, son CUI, son type sémantique et sa définition. Enfin, nous stockons les résultats obtenus à partir de la liste afin d'afficher de manière exhaustive tous les détails relatifs à chaque terme médical dans un tableau.

Si nous prenons en considération le terme "patient" et le soumettons au site officiel de WordNet, différentes informations sont fournies, incluant les relations sémantiques et lexicales associées au mot, ainsi que les définitions possibles en lien avec le contexte de la phrase. Cette illustration est présentée dans la figure2.13 ci-dessous.


Figure 2.13: Illustration des informations fournies par WordNet pour le terme "patient"[7].

Comme illustre la figure précedente, WordNet peut en effet utiliser des informations sur la partie du discours (part of speech) d'un mot pour restreindre les sens possibles dans un contexte donné. Il classe les mots en différentes catégories grammaticales telles que les noms, les verbes, les adjectifs, les adverbes, etc.

En spécifiant la partie du discours d'un mot dans une phrase, WordNet peut fournir une liste restreinte de synsets correspondant à cette catégorie grammaticale spécifique. Par exemple, si le mot en question est Patient comme dans l'exemple, et qu'il est utilisé comme un nom, WordNet ne fournira que les synsets associés à la partie du discours "Noun" pour ce mot.

Cependant, lorsqu'un mot a plusieurs sens possibles dans une catégorie donnée, WordNet se base sur les informations lexicales disponibles pour chaque sens pour effectuer une

[7]http://wordnetweb.princeton.edu/perl/webwn

comparaison avec le contexte spécifique de la phrase. Les caractéristiques lexicales prises en compte peuvent inclure les définitions, les exemples d'utilisation, les relations sémantiques, les synonymes, les antonymes, etc., associés à chaque sens.

## 2.8 Évaluation du Modèle

Le facteur crucial dans cette problématique d'apprentissage automatique réside dans la performance de notre modèle. Pour évaluer les performances de chaque solution de ce modèle, nous pouvons utiliser les mesures suivantes.

Pour évaluer le fine-tuning du modèle BioGPT dans la première solution, nous avons utilisé la perplexité comme métrique. Pour cela, nous avons obtenu les valeurs de la perte d'entraînement (training loss) et de la perte d'évaluation (evaluation loss), où:

- **Perplexité** est une mesure statistique de la confiance avec laquelle un modèle de langage prédit un échantillon de texte. En d'autres termes, elle quantifie le degré de "surprise" du modèle lorsqu'il rencontre de nouvelles données. Plus la perplexité est faible, meilleure est la capacité du modèle à prédire le texte[60].

  La perplexité (PPL) est calculée en utilisant la fonction exponentielle de la perte d'évaluation moyenne du modèle. Mathématiquement, cela peut être exprimé comme suit:

$$PPL = exp(evaluationloss)$$

Dans le contexte de la deuxième solution, afin d'évaluer la performance du résumé de texte par rapport à son texte original, nous avons adopté une approche métrique plus spécifique: les variantes de la métrique ROUGE (Recall-Oriented Understudy for Gisting Evaluation) à savoir ROUGE-1, ROUGE-2 et ROUGE-L. Ces mesures évaluent la similarité entre le résumé généré et le texte de référence en se basant sur les concepts de rappel (Recall), de précision (Precision) et du F1-score.

- **ROUGE-N** est une métrique d'évaluation automatique largement utilisée pour mesurer la similarité entre un texte généré par un modèle et une référence du texte original. N fait référence à la taille des n-grammes, qui sont des séquences de N mots consécutifs dans le texte. ROUGE-N mesure le nombre de n-grammes communs entre le texte généré et la référence, ce qui permet d'évaluer la qualité et la pertinence du texte généré par rapport à l'original[61].

  - **ROUGE-1** également appelé ROUGE-1gram, est une variante spécifique de la métrique ROUGE qui se concentre sur les unigrammes, c'est-à-dire des séquences de mots individuels[61].

* **Précision ROUGE-1** peut être calculée comme le rapport entre le nombre de mots unigrammes dans le résumé généré qui apparaissent également dans le texte de référence (NBGR), et le nombre total de mots unigrammes dans le résumé généré (NBTG).

$$\underset{precision}{ROUGE-1} = NBGR/NBTG$$

* **Rappel ROUGE-1** est calculé en prenant le rapport entre le nombre de mots unigrammes dans le texte de référence qui apparaissent également dans le résumé généré (NBRG) et le nombre total de mots unigrammes dans la référence (NBTR).

$$\underset{recall}{ROUGE-1} = NBRG/NBTR$$

* **F1-score ROUGE-1** peut être directement obtenu à partir de la précision ROUGE-1 et du rappel ROUGE-1 en utilisant la formule standard du score F1.
  Le score F1 est une mesure de la qualité globale d'un système d'évaluation. Il est calculé en utilisant à la fois la précision et le rappel[62].
  La formule du score F1 est la suivante :

$$F1-score = 2*(precision*recall)/(precision+recall)$$

– **ROUGE-2** la procédure est similaire à celle de ROUGE-1, mais en se concentrant sur les bigrammes plutôt que les unigrammes[61].
  Voici comment calculer la précision ROUGE-2, le rappel ROUGE-2 et le score F1 ROUGE-2:

  * **Précision ROUGE-2** mesure le rapport entre le nombre de bigrammes dans le résumé généré qui apparaissent également dans le texte de référence (NBGR), et le nombre total de bigrammes dans le résumé généré (NBTG).

$$\underset{precision}{ROUGE-2} = NBGR/NBTG$$

  * **Rappel ROUGE-2** est calculé en prenant le rapport entre le nombre de bigrammes dans le texte de référence qui apparaissent également dans le résumé généré (NBRG), et le nombre total de bigrammes dans la référence (NBTR).

$$\underset{recall}{ROUGE-2} = NBRG/NBTR$$

  * **F1-score ROUGE-2** peut être obtenu directement en utilisant la formule standard du score F1, en utilisant la précision ROUGE-2 et le rappel ROUGE-2. Cette formule est similaire à celle utilisée pour le score F1 ROUGE-1.

– **ROUGE-L** est basé sur la plus longue sous-séquence commune (LSC) entre la sortie de notre modèle et la référence, c'est-à-dire la séquence la plus longue de mots (non nécessairement consécutifs, mais toujours dans l'ordre) partagée entre les deux. Une séquence partagée plus longue devrait indiquer une similarité plus importante entre les deux séquences.[61].

  Nous pouvons calculer le rappel, la précision et le score F1 ROUGE-L de la même manière que nous l'avons fait avec ROUGE-N, mais cette fois nous remplaçons chaque correspondance de n-gramme par la LSC.

  * **Précision ROUGE-L** est le rapport entre la longueur de la plus longue sous-séquence commune (LSC) et le nombre de mots unigrammes dans le résumé générée (NBG).

$$\underset{precision}{ROUGE - L} = LSC/NBG$$

  * **Rappel ROUGE-L** est en effet le rapport entre la longueur de la plus longue sous-séquence commune (LSC) et le nombre de mots unigrammes dans la référence (NBR).

$$\underset{recall}{ROUGE - L} = LSC/NBR$$

  * **F1-score ROUGE-L** peut être directement obtenu à partir de la précision ROUGE-L et du rappel ROUGE-L en utilisant la formule standard du score F1. Cette formule est similaire à celle utilisée pour le calcul du score F1 de ROUGE-N.

## 2.9 Conclusion

Ce chapitre a constitué une représentation conceptuelle complète des différents éléments présents dans notre travail. Nous avons abordé l'ensemble des données utilisées, les étapes de pré-traitement nécessaires, les traitements de textes utilisés, l'analyse sémantique mise en œuvre, ainsi que l'évaluation de notre travail. Dans le chapitre suivant, nous passerons à l'implémentation pratique de ces concepts afin de créer un modèle opérationnel prêt à être utilisé.

# Chapitre 3

# Implémentation

## 3.1 Introduction

Ce chapitre aborde la mise en œuvre des concepts décrits dans le chapitre précédent et offre une vue d'ensemble détaillée de l'ensemble du processus. Il comprendra la présentation de l'environnement utilisé, ainsi que des bibliothèques et des API employées. De plus, il exposera l'implémentation en fournissant des détails de code, les résultats de la génération de texte et ses domaines d'application. Enfin, dans la partie simulation, nous observerons les résultats en action grâce à une interface conviviale pour l'utilisateur.

## 3.2 Ressources Matérielles

Le tableau ci-après3.1 décrit les spécifications de l'ordinateur qui a été utilisé pour les tests et l'entraînement.

| | |
|---|---|
| CPU | Intel i5-8250U 8th generation 1.80 GHz |
| GPU | UHD Graphics 620 |
| Stockage | 256 SSD |
| Mémoire | 8 Go DDR4 |
| Operating System | Windows 10 Professionnel 21H2 64 bits |

Table 3.1: Spécifications de l'ordinateur utilisé.

## 3.3 Ressources Logicielles

Dans cette section, nous présenterons les ressources logicielles utilisées lors de la mise en œuvre, notamment le langage de programmation, les frameworks ainsi que les bibliothèques.

### 3.3.1 Environnement de Travail

Dans l'environnement de développement, nous avons utilisé

- **Google Colaboratory** un IDE Python basé sur navigateur qui permet à quiconque d'écrire et d'exécuter du code Python arbitraire, ce qui est particulièrement utile pour l'apprentissage automatique et l'analyse de données. Il offre des ressources informatiques gratuites avec une accélération de l'entraînement GPU[1].
- **JupyterLab** est un environnement de développement interactif. Il est basé sur le web et fournit une plate-forme flexible et puissante pour l'analyse de données, l'apprentissage automatique et prend en charge plusieurs langages de programmation[2].

### 3.3.2 Bibliothèques

**Python** comme langage de programmation, un langage de haut niveau qui contient de nombreuses bibliothèques qui ont facilité le traitement de nos données et la construction de nos modèles d'apprentissage profond.

Nous avons utilisé les bibliothèques suivantes :

- **Pandas** est une bibliothèque logicielle écrite pour le langage de programmation Python, utilisée pour la manipulation et l'analyse de données[3].
- **Numpy** est une bibliothèque pour le langage de programmation Python, qui ajoute la prise en charge de tableaux et de matrices de grande dimension, ainsi qu'une vaste collection de fonctions mathématiques de haut niveau pour opérer sur ces tableaux[4].
- **Matplotlib** est une bibliothèque logicielle écrite pour le langage de programmation Python, utilisée pour la manipulation et l'analyse de données, ainsi que pour la création de graphiques et de visualisations[5].
- **Spacy** est une bibliothèque logicielle pour le langage de programmation Python. Elle offre des fonctionnalités avancées telles que l'analyse syntaxique, la reconnaissance d'entités nommées, le marquage morphologique et d'autres étapes clés dans le prétraitement des données textuelles[6].
- **Requests** est une bibliothèque Python permettant d'envoyer des requêtes HTTP de manière simple et efficace, utilisée pour interagir avec des API web et récupérer des données à partir de ressources en ligne[7].

[1] https://colab.research.google.com/
[2] https://jupyter.org/
[3] https://pandas.pydata.org/
[4] https://numpy.org/
[5] https://matplotlib.org/
[6] https://spacy.io/
[7] https://pypi.org/project/requests/

- **Sumy** est une bibliothèque Python utilisée pour la génération automatique de résumés de texte. Elle offre plusieurs algorithmes de résumé, tels que LSA (Latent Semantic Analysis), LexRank et Luhn, permettant de condenser un texte en un résumé concis et informatif[8].

- **TensorFlow** est une bibliothèque logicielle développée pour le langage de programmation Python, destinée à la manipulation et à l'analyse de données[9].

- **NLTK** (Natural Language Toolkit) est une bibliothèque Python largement utilisée. Elle offre de nombreuses fonctionnalités pour la manipulation et l'analyse de texte, telles que la tokenisation, la lemmatisation, la classification et l'analyse syntaxique. NLTK propose également des ressources linguistiques pour faciliter le développement de projets de TALN[10].

- **Scikit-learn** est une bibliothèque logicielle de machine learning gratuite pour le langage de programmation Python. Elle propose divers algorithmes de classification, de régression et de regroupement (clustering) pour l'analyse et la modélisation de données[11].

### 3.3.3 Outils de Gestion

Dans notre projet, nous avons utilisé des outils de gestion de projet ainsi que des systèmes de contrôle pour assurer une gestion efficace du projet et un suivi des modifications apportées au code source.

- **Notion** est un logiciel de productivité et de collaboration qui permet aux individus et aux équipes de gérer des tâches. Il fournit un espace de travail unifié permettant aux membres de l'équipe de travailler et de partager des informations[12].

- **Git** est un système de contrôle de version distribué gratuit et open source utilisé pour suivre les modifications du code source pendant le développement de logiciels. Il permet à plusieurs développeurs de travailler simultanément sur la même base de code, en gardant une trace des modifications et en les fusionnant au besoin[13].

- **Google Drive** est un service de stockage de fichiers basé développé par Google. Il permet aux utilisateurs de stocker des fichiers sur leurs serveurs, de synchroniser des fichiers sur plusieurs appareils et de partager des fichiers avec d'autres[14].

[8] https://pypi.org/project/sumy/
[9] https://www.tensorflow.org/
[10] https://www.nltk.org/index.html
[11] https://scikit-learn.org/stable/
[12] https://www.notion.so/fr-fr
[13] https://git-scm.com/
[14] https://drive.google.com/

## 3.4 Ensemble de Données

À l'aide de la bibliothèque JSON, nous extrayons les fichiers au format JSON et les convertissons en format TXT à l'aide du code suivant3.1. Cette opération est effectuée pour les ensembles de données d'entraînement, de test et de validation.

```
import json
import os

with open('C:/Users/21367/Desktop/PFE/Datasets/cochrane_english/test/test.json',
          'r', encoding='utf-8') as f:
    data = json.load(f)
for i, example in enumerate(data):
    if i >= 0:
        source_text = example['source']
        target_text = example['target']
        filename = f'test_{i}.txt'
        with open(os.path.join('C:/Users/21367/Desktop/PFE/Datasets/cochrane_english/test', filename),
                  'w', encoding='utf-8') as f:
            f.write(source_text + '\n')
            f.write(target_text + '\n')
```

Figure 3.1: Code pour faire extraire et convertir les données de Cochrane.

## 3.5 Pré-traitement des Données

L'étape de prétraitement revêt une importance capitale dans l'amélioration de la qualité des données, notamment lorsqu'il s'agit de l'apprentissage du modèle. La tokenisation est réalisée à l'aide de la bibliothèque nltk en téléchargeant le module "punkt" pour segmenter les phrases et les mots.

La phase de nettoyage des données peut être réalisée de différentes manières, mais dans ce travail, nous utilisons la bibliothèque nltk pour téléchager la liste des mots vides (stopwords) en anglais, ainsi que le module "wordnet" pour la lemmatisation. Nous faisons également appel aux modules "string" et "re" pour traiter les ponctuations et les caractères inutiles.

Ci-dessous est présenté un tableau3.2 illustrant un exemple de texte médical avant le pré-traitement accompagné de sa version après le pré-traitement.

| Texte d'entrée avant | Texte après |
|---|---|
| One trial involving 92 participants with pretreated/relapsed WM compared the effect of fludarabine versus the combination of cyclophosphamide (the alkylating agent), doxorubicin and prednisone (CAP). Compared to CAP, the Hazard ratio (HR) for deaths of treatment with fludarabine was estimated to be 1.04, with a standard error of 0.30 (95 % CI 0.58 to 1.48) and it indicated that the mean difference of median survival time was -4.00 months, and 16.00 months for response duration. The relative risks (RR) of response rate was 2.80 (95 % CI 1.10 to 7.12). There were no statistically difference in overall survival rate and median survival months, while on the basis of response rate and response duration, fludarabine seemed to be superior to CAP for pretreated/relapsed patients with macroglobulinaemia. Although alkylating agents have been used for decades they have never actually been tested in a proper randomised trial. This review demonstrated that there is currently no evidence to suggest that alkylating agents are effective in treating Waldenstrom's macroglobulinaemia. | one trial involving 92 participant pretreated/relapsed wm compared effect fludarabine versus combination cyclophosphamide alkylating agent doxorubicin prednisone cap . compared cap hazard ratio hr death treatment fludarabine estimated 1.04 standard error 0.30 95 ci 0.58 1.48 indicated mean difference median survival time -4.00 month 16.00 month response duration . relative risk rr response rate 2.80 95 ci 1.10 7.12 . statistically difference overall survival rate median survival month basis response rate response duration fludarabine seemed superior cap pretreated/relapsed patient macroglobulinaemia . although alkylating agent used decade never actually tested proper randomised trial . review demonstrated currently evidence suggest alkylating agent effective treating waldenstrom 's macroglobulinaemia |

Table 3.2: Exemple de texte avant et après avoir effectuer le pré-traitement.

## 3.6 Extraction des Entités Nommées

Le modèle "en-ner-bionlp13cg-md" est un modèle NER spaCy entraîné sur le corpus BIONLP13CG, il est spécifiquement conçu pour extraire des entités biomédicales à partir de données textuelles. Il utilise une combinaison d'approches basées sur des règles et sur l'apprentissage automatique pour identifier divers types d'entités que l'on trouve couramment dans la littérature biomédicale, telles que CANCER, CELL, ORGAN, ORGANISM, SIMPLE-CHEMICAL, TISSUE..etc, la figure suivante3.2 présente les types d'entités de ce modèle et sa performance de F1 score à 76.57 des étiqueteurs de parties du discours biomédicaux à la pointe.

| model | F1 | Entity Types |
|---|---|---|
| en_ner_bionlp13cg_md | 76.57 | AMINO_ACID, ANATOMICAL_SYSTEM, CANCER, CELL, CELLULAR_COMPONENT, DEVELOPING_ANATOMICAL_STRUCTURE, GENE_OR_GENE_PRODUCT, IMMATERIAL_ANATOMICAL_ENTITY, MULTI-TISSUE_STRUCTURE, ORGAN, ORGANISM, ORGANISM_SUBDIVISION, ORGANISM_SUBSTANCE, PATHOLOGICAL_FORMATION, SIMPLE_CHEMICAL, TISSUE |

Figure 3.2: Représentation des types des entités et du F1 score du modèle bionlp13cg.

Pour effectuer la reconnaissance d'entité nommée (NER) à l'aide du modèle en-ner-bionlp13cg-md de la bibliothèque spaCy il faut d'abord charger le modèle pour passer l'ensemble de fichiers texte déjà prétraité. Après le traitement de chaque fichier, les entités reconnues sont extraites et enregistrées dans de nouveaux fichiers dans un répertoire spécifié. Au final, on obtient les fichiers de sortie qui contiennent une entité reconnue par ligne, chaque ligne étant constituée du texte de l'entité et de son étiquette, séparés par une tabulation comme indiqué ci-dessous3.3:

```
people  ORGANISM
arterial        MULTI_TISSUE_STRUCTURE
ulcer   CANCER
blood concentrated growth factor        ORGANISM_SUBSTANCE
arterial leg    PATHOLOGICAL_FORMATION
arterial ulcer  PATHOLOGICAL_FORMATION
human   ORGANISM
nonhealing ulcer        CANCER
wound   PATHOLOGICAL_FORMATION
ketanserin      SIMPLE_CHEMICAL
arterial ulcer  CANCER
ulcer venous    MULTI_TISSUE_STRUCTURE
bloodderived concentrated growth factor GENE_OR_GENE_PRODUCT
patient ORGANISM
```

Figure 3.3: Représentation de l'entité et de son étiquette.

Les fichiers produits lors de l'étape NER seront ensuite employés pour annoter l'ensemble de données Cochrane lors de la phase de fine-tuning du modèle pré-entrainé BioGPT, que nous examinerons ultérieurement.

## 3.7 Génération de Texte

Afin d'accomplir cette étape, nous avons procédé au finetuning du modèle pré-entraîné BioGPT de la manière suivante :

### 3.7.1 Fine-tuning BioGPT

BioGPT[15]est un modèle de langage pré-entraîné, il est développé par Microsoft qui est spécifiquement conçu pour les tâches biomédicales de traitement du langage naturel (NLP).

Dans ce qui suite nous présentons les étapes de l'affinement de BioGPT:

- **Préparation des données:** constitue la première étape, au cours de laquelle nous avons traité et annoté notre ensemble de données (entraînement et validation) pour qu'il soit conforme au format d'entrée requis par le modèle BioGPT. Ces données pré-traitées ont été étiquetées en fonction des entités extraites lors de l'étape précédente, garantissant ainsi leur compatibilité avec le modèle BioGPT.
- **Chargement du modèle pré-entraîné:** dans cette étape, nous avons chargé le modèle BioGPT pré-entraîné en mémoire ainsi que son tokenizer à l'aide d'un cadre d'apprentissage automatique tel que TensorFlow comme illustre la figure 3.4 suivante.

```
model_name = "microsoft/biogpt"
tokenizer = BioGptTokenizer.from_pretrained(model_name)
model = BioGptForCausalLM.from_pretrained(model_name)
```

Figure 3.4: Chargement du modèle BioGPT.

- **Tokenisation des données:** cette étape est faite à l'aide de la méthode tokenizer.encode() de la bibliothèque Transformers pour transformer les données textuelles d'entraînement et de validation en séquences de tokens, en tenant compte des paramètres de troncation et de la longueur maximale définis. Dans la figure 3.5 suivante on montre cette étape.

```
train_tokenized_data = [tokenizer.encode(text.strip(), truncation=True, max_length=1024) for text in train_data]
valid_tokenized_data = [tokenizer.encode(text.strip(), truncation=True, max_length=1024) for text in validation_data]
```

Figure 3.5: Tokenisation des données.

- **Définition des arguments de configuration pour l'entraînement du modèle:** L'utilisation de la classe TrainingArguments a été privilégiée, celle-ci étant dédiée à la configuration des paramètres d'entraînement d'un modèle. Elle permet de définir les multiples options et paramètres associés à cette phase de l'apprentissage. La figure 3.6 suivante illustre les détails de cette étape.

[15]https://github.com/microsoft/BioGPT

```
training_args = TrainingArguments(
    output_dir="/results",  # Répertoire de sortie pour les résultats de l'entraînement
    overwrite_output_dir=True,  # Pour écraser le répertoire de sortie s'il existe déjà
    num_train_epochs=8,  # Nombre d'époques d'entraînement
    per_device_train_batch_size=16,  # Taille du lot d'entraînement
    per_device_eval_batch_size=16,  # Taille du lot d'évaluation
    eval_steps=500,  # Fréquence d'évaluation pendant l'entraînement (en nombre de pas)
    save_steps=2000,  # Fréquence d'enregistrement des points de contrôle (en nombre de pas)
    warmup_steps=500,  # Nombre de pas d'échauffement pour ajuster le taux d'apprentissage
    learning_rate=2e-5,  # Taux d'apprentissage initial pour l'optimiseur
    logging_dir="/logs",  # Répertoire où les journaux de suivi seront enregistrés
    logging_steps=500,  # Fréquence d'enregistrement des journaux de suivi (en nombre de pas)
    evaluation_strategy="epoch",  # Pour l'évaluation à chaque époque
    do_train=True,  # Indicateur pour effectuer l'entraînement
    do_eval=True,  # Indicateur pour effectuer l'évaluation
)
```

Figure 3.6: Arguments de configuration pour l'entraînement du modèle.

Pour récapituler, ces paramètres incluent le répertoire de sortie, le nombre d'époques, la taille des lots, la fréquence d'évaluation et d'enregistrement des points de contrôle, ainsi que le taux d'apprentissage, entre autres, permettent de configurer les divers paramètres et options associés à l'entraînement de notre modèle.

## 3.8 Résumé de Texte

Tel qu'indiqué précédemment dans la section de conception, notre travail se base sur l'une des approches distinctes pour la génération de résumés textuels: l'approche extractive et l'approche abstractive. Pour cette étude, nous avons choisi d'utiliser l'approche extractive, qui implique l'identification des phrases les plus pertinentes du texte en fonction de leur importance, afin de générer un résumé sans créer de nouvelles phrases.

Le pseudo-code de l'image3.7 suivante explique la réalisation de résumés en utilisant l'algorithme TextRank.

```
Fonction summarizeText(texte, numPhrases)
DEBUT
    analyseur = PlaintextParser(texte, Tokenizer("english"));
    summarizer = TextRankSummarizer();

    phrasesRecapitulatives = summarizer(analyseur.document, numPhrases);
    texteRecapitulatif = "";

    pour chaque phrase dans phrasesRecapitulatives faire
        texteRecapitulatif = texteRecapitulatif + phrase;
    fin pour

    RETOURNER texteRecapitulatif
Fin Fonction
```

Figure 3.7: Illustration de résumé un texte en utilisant l'algorithme TextRank à travers un pseudo-code.

La figure3.8 cj-dessous représente le texte original utilisé dans cette étude provient de notre ensemble de données Cochrane.

Two trials met the inclusion criteria. One compared 2 ketanserin ointment in polyethylene glycol (PEG) with PEG alone, used twice a day by 40 participants with arterial leg ulcers, for eight weeks or until healing, whichever was sooner. One compared topical application of blood-derived concentrated growth factor (CGF) with standard dressing (polyurethane film or foam); both applied weekly for six weeks by 61 participants with non-healing ulcers (venous, diabetic arterial, neuropathic, traumatic, or vasculitic). Both trials were small, reported results inadequately, and were of low methodological quality. Short follow-up times (six and eight weeks) meant it would be difficult to capture sufficient healing events to allow us to make comparisons between treatments. One trial demonstrated accelerated wound healing in the ketanserin group compared with the control group. In the trial that compared CGF with standard dressings, the number of participants with diabetic arterial ulcers were only reported in the CGF group (9/31), and the number of participants with diabetic arterial ulcers and their data were not reported separately for the standard dressing group. In the CGF group, 66.6 (6/9) of diabetic arterial ulcers showed more than a 50 decrease in ulcer size compared to 6.7 (2/30) of non-healing ulcers treated with standard dressing. We assessed this as very-low certainty evidence due to the small number of studies and arterial ulcer participants, inadequate reporting of methodology and data, and short follow-up period ..... there were only two studies and few participants with arterial disease, and because the studies were short and reported few results. This made it impossible to determine whether there was any real difference in the number of ulcers healed between the groups.

Figure 3.8: Texte original utilisé.

Nous présentons dans la figure3.9 suivante également le résumé généré à l'aide de

l'algorithme TextRank.

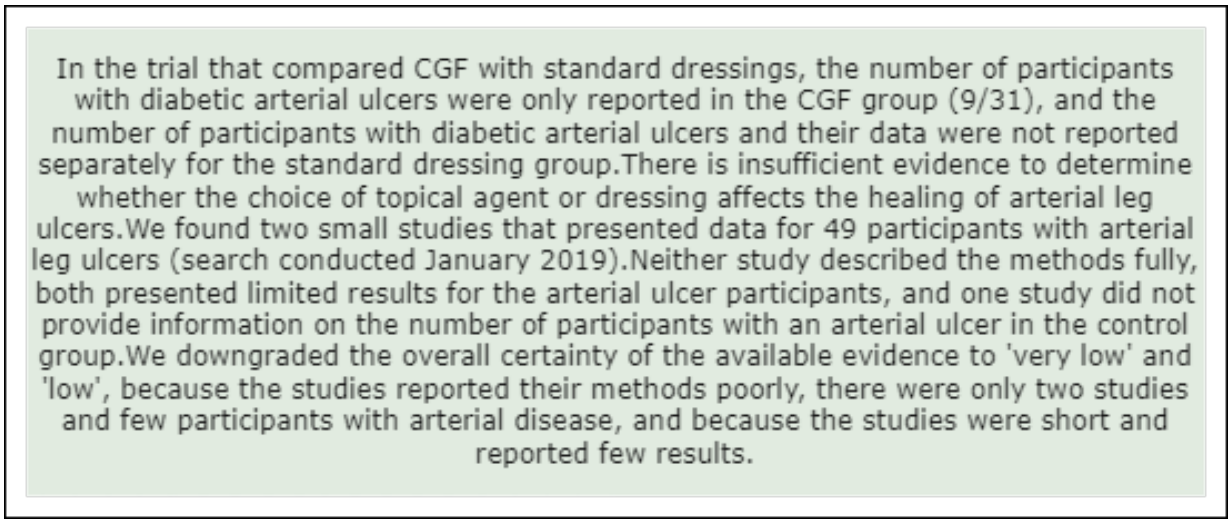


Figure 3.9: Résumé généré à l'aide de l'algorithme TextRank.

## 3.9 Simplification de Texte

Le processus de simplification peut être décomposé en plusieurs étapes, incluant notamment l'analyse sémantique en utilisation les ontologies et les ressources sémantiques cités dans la conception, telles que :

L'étape initiale consiste à extraire les termes médicaux susceptibles de poser des difficultés de compréhension. Pour réaliser cette étape, nous avons utilisé deux modèles, à savoir "en-ner-bc5cdr-md" et "en-ner-bionlp13cg-md" de spaCy, pour extraire les termes médicaux du texte. En parcourant les entités détectées par les modèles et vérifie si l'étiquette d'entité correspond à l'une des étiquettes spécifiées pour les termes médicaux.

Par la suite, nous procédons à l'obtention des informations détaillées concernant les termes médicaux. Nous avons effectué pour chaque terme une recherche en utilisant l'API des services de terminologie de l'UMLS afin de récupérer des détails spécifiques aux termes médicaux.

Dans un premier temps, nous tentons de trouver l'identificateur unique de concept (CUI) en effectuant une requête à l'API en utilisant le terme. Une fois le terme est trouvé, nous récupérons l'identificateur associé à ce terme.

Une fois que l'identificateur unique de concept (CUI) est obtenu, le type sémantique du terme est récupéré en effectuant une requête à l'API. Par la suite, une autre requête a été employée pour extraire la définition à partir de l'identificateur unique de concept (CUI) obtenu. Les définitions ont été filtrées en fonction de sources racine spécifiques, à savoir MSH, SNOMED-US, HPO et MEDLINEPLUS.

Par la suite, un dictionnaire a été élaboré, comprenant le terme, le CUI, le type sémantique et la définition associés. Ce dictionnaire a été ajouté à la liste des résultats.

En conclusion, un DataFrame Pandas a été créé à partir de la liste des résultats afin de présenter de manière tabulaire tous les détails associés à chaque terme médical.

Et afin d'assurer la simplification des textes, nous envisageons d'enrichir le contenu en utilisant les définitions fournies par WordNet. Pour ce faire, nous ferons appel aux synsets et à l'algorithme de Lesk préétablis dans WordNet afin d'extraire les définitions d'un mot dans son contexte approprié.

La figure3.10 ci-dessous présente un exemple de synsets, accompagnés de leurs définitions respectives, ainsi que la définition choisie à la fin de chaque mot, en adéquation avec le contexte de la phrase suivante : *"it is hoped that future iterations of this review will benefit from larger sample sizes across a wider geographical area."*

```
--------------------------------------------
Word: iteration
Synset('iteration.n.01') (computer science) a single execution of a set of instructions that are to be repeated
Synset('iteration.n.02') (computer science) executing the same set of instructions a given number of times or until a specified result is obtained
Synset('iteration.n.03') doing or saying again; a repeated performance
iteration : doing or saying again; a repeated performance
--------------------------------------------

Word: area
Synset('sphere.n.01') a particular environment or walk of life
Synset('sphere.n.02') any spherically shaped artifact
Synset('sphere.n.03') the geographical area in which one nation is very influential
Synset('sector.n.03') a particular aspect of life or activity
Synset('sphere.n.05') a solid figure bounded by a spherical surface (including the space it encloses)
Synset('sphere.n.06') a three-dimensional closed surface such that every point on the surface is equidistant from the center
Synset('celestial_sphere.n.01') the apparent surface of the imaginary sphere on which celestial bodies appear to be projected
area : a particular environment or walk of life
--------------------------------------------
```

Figure 3.10: Exemple de synsets ainsi que leurs définitions appropriées.

## 3.10 Évaluation du Modèle

### 3.10.1 Évaluation de la Première Solution

Dans le cadre de notre processus d'entraînement du modèle, visant à obtenir un modèle BioGPT finetuné spécifiquement sur notre ensemble de données de 3 560 fichiers d'entraînement et évalué sa performance sur 411 fichiers de validation, nous avons calculé les pertes d'entraînement (training loss) et les pertes d'évaluation (eval loss) comme métriques clés pour évaluer la qualité de notre modèle. Les pertes d'entraînement mesurent l'erreur sur l'ensemble des données d'entraînement, tandis que les pertes d'évaluation mesurent l'erreur sur l'ensemble des données de validation.

Nous avons noté les résultats suivants, qui sont présentés dans la figure3.11 suivante, ces valeurs atteint de perte d'entraînement et de perte d'évaluation nous permettent d'évaluer la capacité de notre modèle à s'ajuster aux données d'entraînement ainsi qu'à généraliser ses prédictions sur les données de validation. L'écart entre les pertes d'entraînement et les pertes d'évaluation nous donne une indication de la capacité de notre modèle à éviter le surajustement (overfitting) et à généraliser de manière efficace.

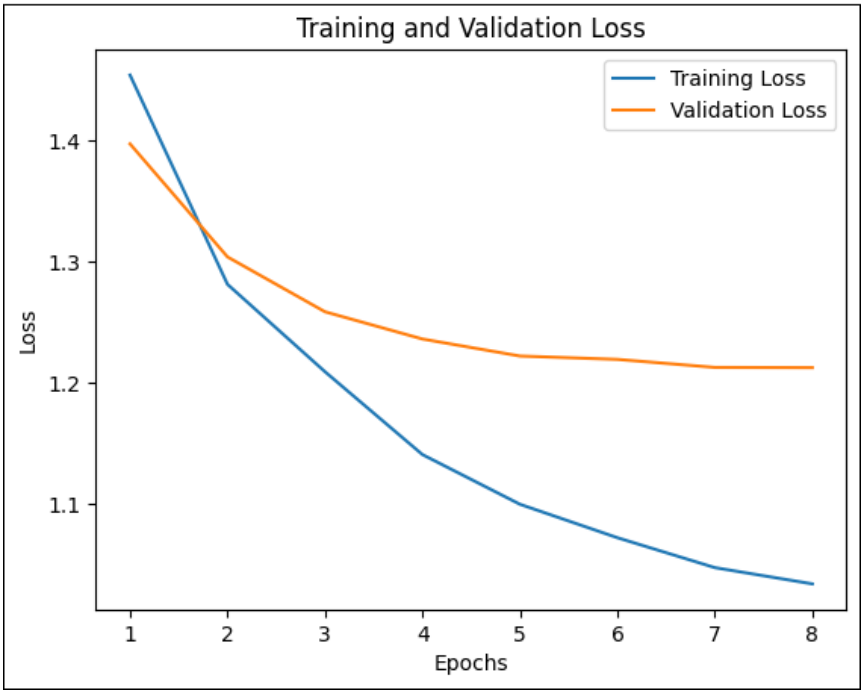


Figure 3.11: Pertes d'entraînement et pertes d'évaluation lors de l'entraînement du modèle.

En utilisant ces métriques de pertes, nous sommes en mesure de quantifier la performance de notre modèle et de suivre sa progression tout au long de l'entraînement.

Nous avons aussi calculé la perplexité à chaque époque. La perplexité, en tant que métrique, sert à évaluer la qualité de prédiction du modèle sur un ensemble de données. Plus la perplexité est faible, meilleure est la capacité du modèle à prédire les données de test.

Nous avons représenté graphiquement l'évolution de la perplexité au fil des époques dans la figure3.12 ci-dessous. Cette visualisation permet d'observer comment la perplexité diminue progressivement à mesure que le modèle apprend et s'améliore dans sa prédiction des données.

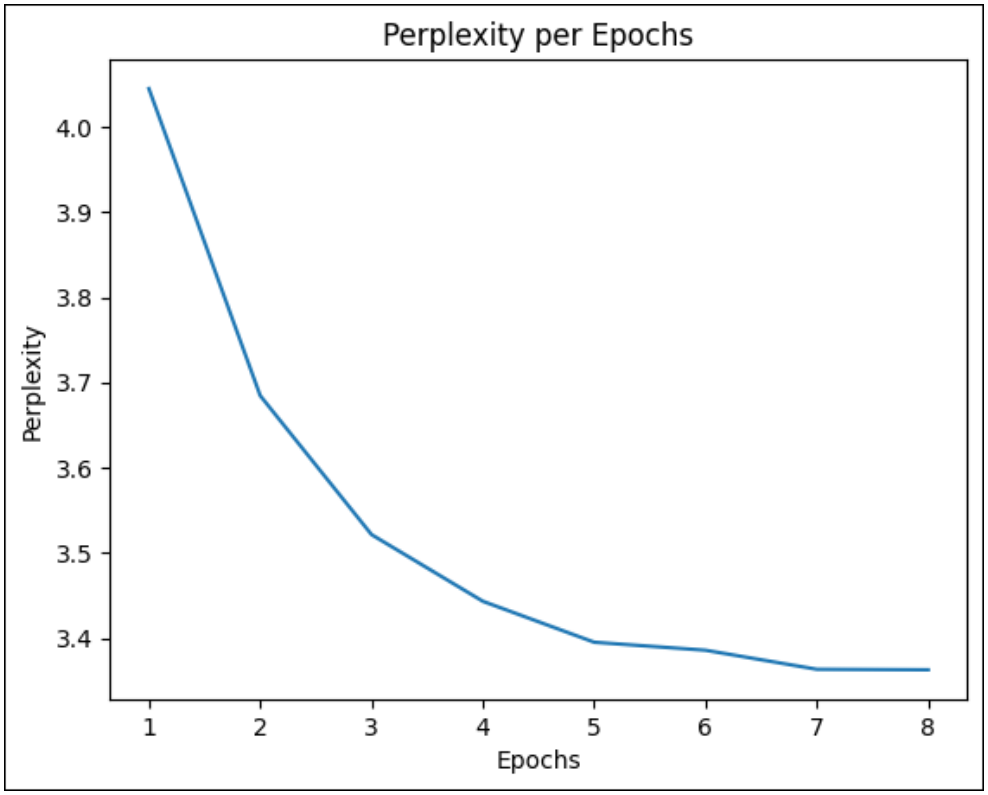


Figure 3.12: Perplexité par époque lors de l'entraînement du modèle BioGPT.

### 3.10.2 Évaluation de la Deuxième Solution

Dans le contexte de la deuxième solution, nous avons adopté une approche différente pour évaluer la performance du résumé de texte par rapport à son texte original. Nous avons utilisé les variantes de la métrique ROUGE (Recall-Oriented Understudy for Gisting Evaluation), notamment ROUGE-1, ROUGE-2 et ROUGE-L. Ces mesures évaluent la similarité entre le résumé généré et le texte de référence en se basant sur les concepts de rappel, de précision et de score F1.

Les résultats obtenus pour les métriques ROUGE sont les suivants:

| **Mesure ROUGE** | **Précision** | **Rappel** | **F1 Score** |
|---|---|---|---|
| ROUGE-1 | 1 | 0.322957 | 0.488235 |
| ROUGE-2 | 0.978417 | 0.283333 | 0.439418 |
| ROUGE-L | 1 | 0.322957 | 0.488235 |

Table 3.3: Résultats de la similarité des résumés générés avec ROUGE.

Ces résultats nous donnent une indication de la similarité entre le résumé généré et le texte de référence, en se concentrant sur les unigrammes (ROUGE-1), les bigrammes (ROUGE-2) et la plus longue sous-séquence commune (ROUGE-L). Une précision élevée

indique que le résumé généré contient un grand nombre de mots présents dans le texte de référence, tandis qu'un rappel élevé indique que le résumé généré couvre une grande partie du texte de référence. Le score F1 combine ces deux mesures en un seul nombre pour évaluer la qualité globale du résumé.

Dans notre cas, les scores de précision, de rappel et de F1 sont généralement élevés, ce qui suggère que le résumé généré est assez fidèle au texte de référence. Cependant, il est important de noter que la précision ROUGE-1 est de 1, ce qui signifie que tous les mots unigrammes dans le résumé généré sont également présents dans le texte de référence. Cela peut indiquer une certaine redondance ou une restriction du vocabulaire dans le résumé généré.

Toutefois, il convient de souligner que la précision ROUGE-2 atteint 0,978417, révélant ainsi que la grande majorité des bigrammes présents dans le résumé généré sont également présents dans le texte de référence, cette constatation suggère une certaine similarité et cohérence au niveau des phrases produites par le modèle. Cependant, le rappel ROUGE-2 se situe à 0,283333, indiquant que le résumé généré ne couvre qu'une fraction réduite des bigrammes présents dans le texte de référence.

En ce qui concerne ROUGE-L, il convient de souligner que la précision atteint la valeur de 1, ce qui implique que la plus longue sous-séquence commune (LSC) entre le résumé généré et le texte de référence correspond à la longueur du résumé généré lui-même. Cette observation suggère une correspondance solide entre des parties spécifiques du texte de référence et le résumé généré. Toutefois, le rappel ROUGE-L affiche une valeur de 0,322957, suggérant ainsi que le résumé généré ne capture qu'une partie limitée de la LSC totale avec le texte de référence.

## 3.11 Déploiement du Modèle

Dans le cadre de cette recherche, nous avons développé une interface utilisant Streamlit.

Streamlit[16]est une bibliothèque Python open-source qui simplifie la création et le partage d'applications web personnalisées et esthétiques dédiées à l'apprentissage automatique et à la science des données.

Tout d'abord, l'utilisateur a la possibilité de choisir entre trois options: effectuer une génération de texte, réaliser un résumé de texte ou simplifier un texte. Cette sélection s'effectue à partir d'une liste comme illustre la figure3.13 suivante, offrant ainsi une flexibilité dans le choix du processus souhaité.

[16]https://docs.streamlit.io/

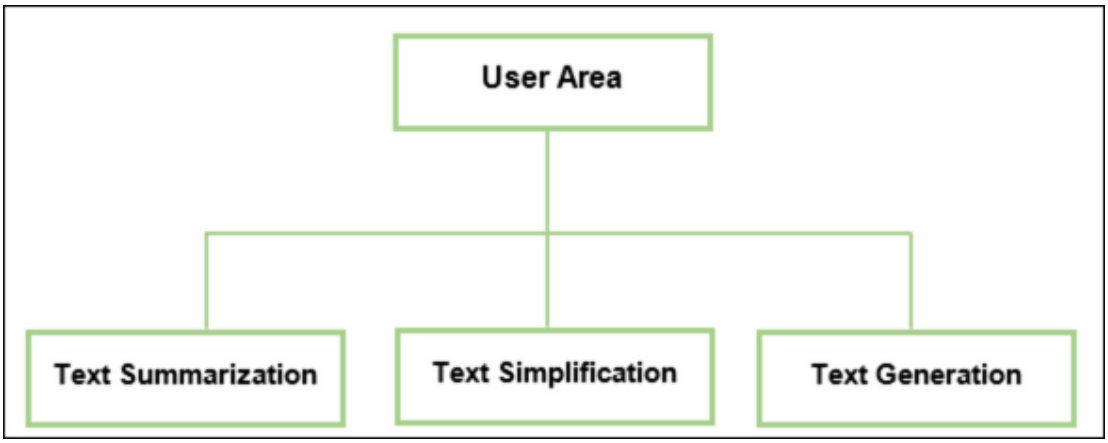


Figure 3.13: Schéma d'accessibilité de notre interface.

### 3.11.1 Génération de Texte

Lorsque l'utilisateur opte pour la génération de texte, l'interface illustrée dans la figure3.14 ci-dessous est présentée. L'utilisateur a alors la possibilité de rédiger une à deux phrases pour effectuer la génération, l'interface intègre également un slider permettant à l'utilisateur de régler la taille maximale souhaitée du texte généré. Ensuite, en cliquant sur le bouton "Generate", le processus de génération est lancé.

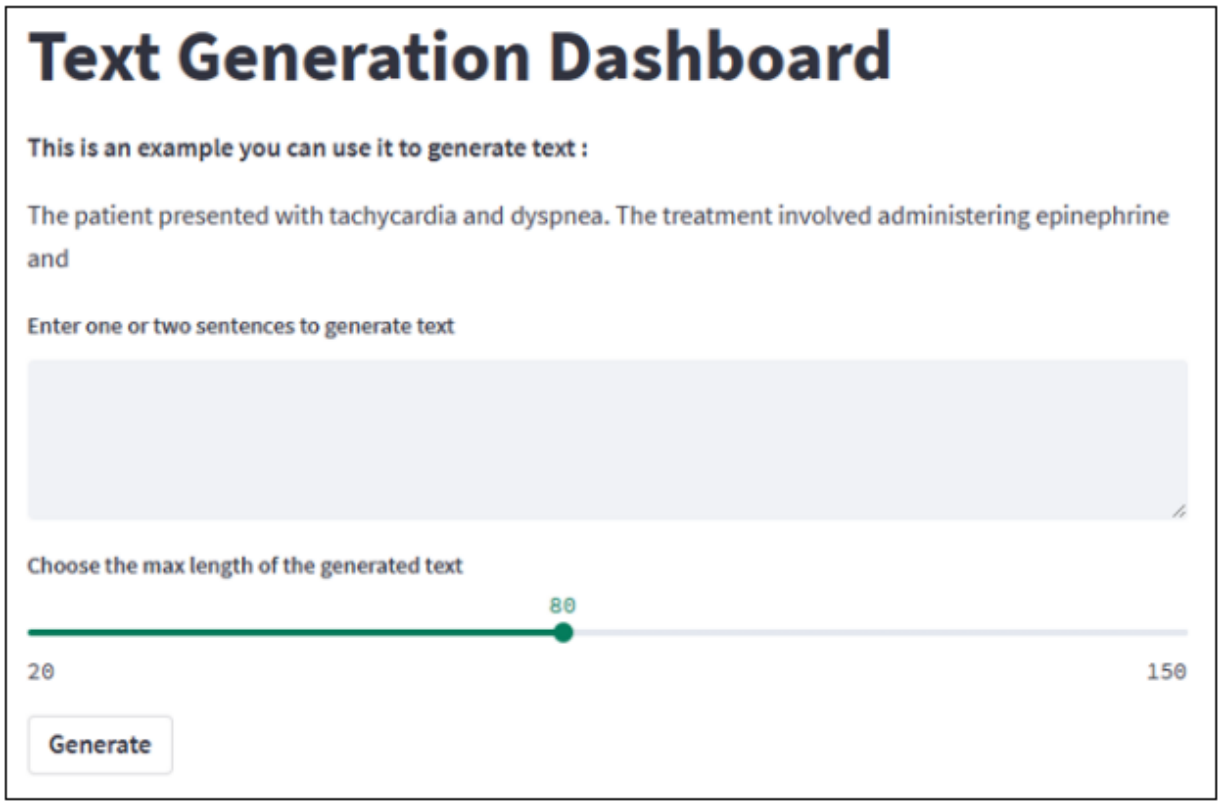


Figure 3.14: Interface de la génération de texte.

Dans cette situation, deux résultats sont obtenus: tout d'abord, un texte généré en fonction de la rédaction de l'utilisateur, où en mettant le curseur sur n'importe quel mot, il est possible d'accéder à une définition de ce mot grâce à WordNet. De plus, un tableau est affiché, contenant les mots médicaux accompagnés de leurs définitions,Comme l'illustre l'image3.15 suivante.

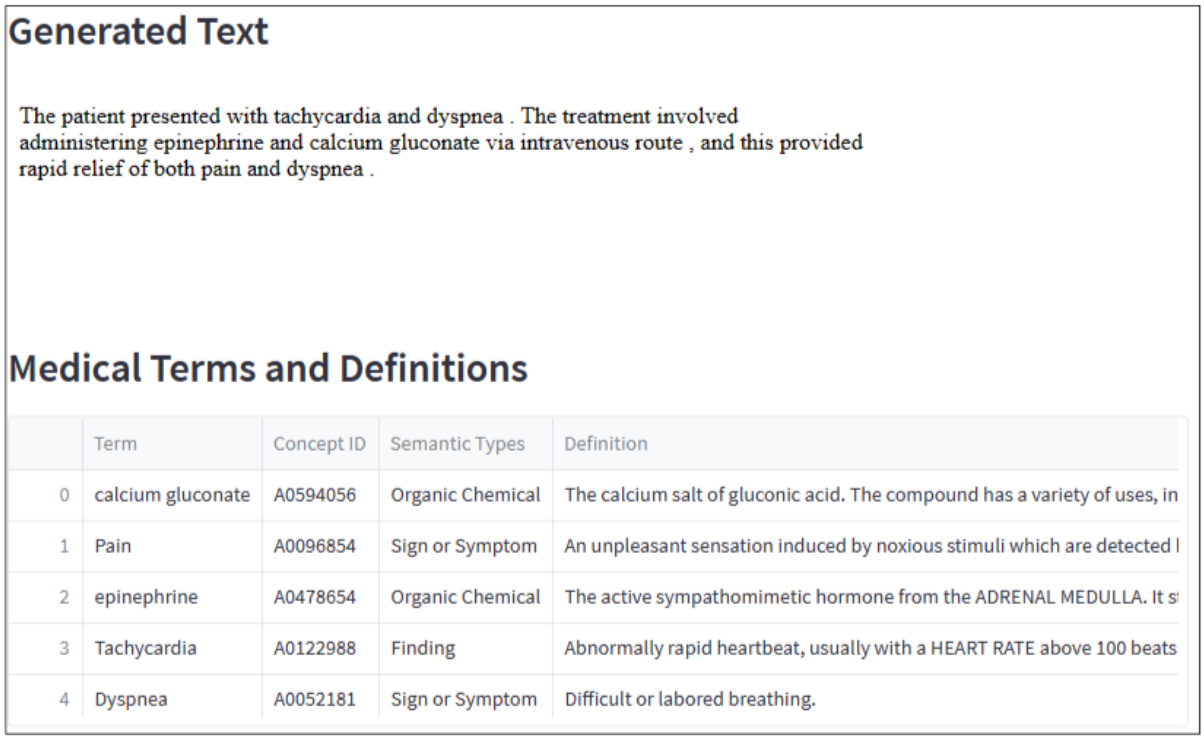


Figure 3.15: Interface du résultats de la génération de texte.

### 3.11.2 Résumé de Texte

Dans le contexte du résumé de texte, l'interface3.16 suivante est présentée à l'utilisateur, lui offrant ainsi la possibilité d'entrer son texte long dans la zone de saisie et de cliquer sur le bouton "Summarize" pour effectuer la phase de résumé.


Figure 3.16: Interface du résumé de texte.

Tel qu'illustré dans la figure3.17 ci-dessous, ce cas se traduit par l'obtention de deux résultats simultanés: tout d'abord, un tableau est présenté, regroupant les termes médicaux du texte original accompagnés de leurs définitions. De plus, un résumé du texte est également affiché, permettant aux utilisateurs de survoler n'importe quel mot et d'accéder à sa définition grâce à WordNet.

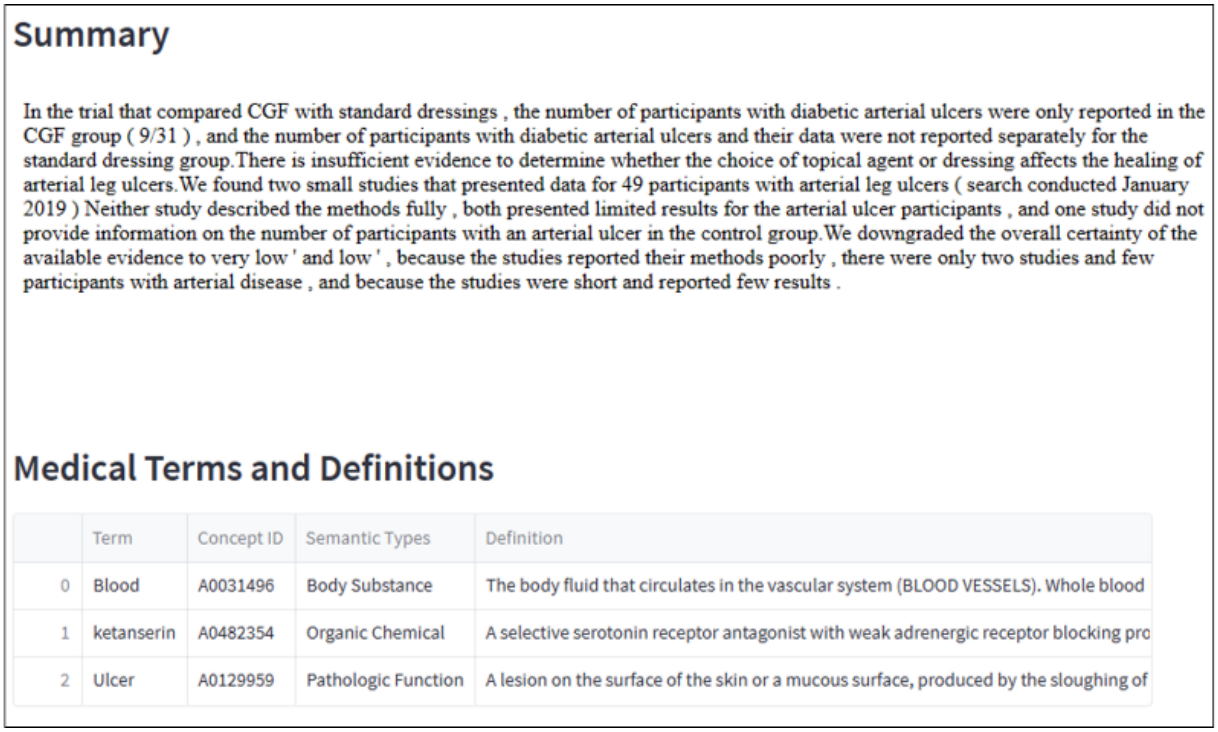


Figure 3.17: Interface du résultats du résumé de texte.

## 3.12 Conclusion

Ce chapitre d'implémentation présente une analyse détaillée de toutes les étapes de conception qui ont été entreprises pour former notre modèle d'apprentissage profond afin d'évaluer ses performances. De plus, pour valider notre travail, nous avons réalisé des tests sur des données réelles en utilisant plusieurs cas d'étude. Ces tests nous ont permis d'examiner les résultats obtenus et de démontrer comment cette recherche peut apporter une contribution significative aux futures études dans ce domaine.

# Conclusion Générale

Ce travail axé sur la simplification des textes médicaux en anglais, présente deux solutions qui ont été examinées en détail.

La première approche consiste à utiliser le modèle BioGPT fine-tuné sur l'ensemble de données Cochrane pour générer du texte médical. Cette méthode permet de produire des textes médicaux tout en maintenant leur contenu informatif et leur précision.

La deuxième solution adoptée dans cette étude de recherche repose sur l'utilisation de l'algorithme TextRank pour résumer les textes médicaux. Cette approche permet de condenser les informations clés des documents médicaux tout en préservant leur cohérence et leur pertinence.

Dans le but de renforcer la simplification des textes médicaux, la troisième solution représentant une analyse sémantique a été intégrée en exploitant les ontologies de l'UMLS et la ressource sémantique WordNet. Cette approche a permis d'accéder aux définitions des termes médicaux ainsi que des termes complexes, contribuant ainsi à une meilleure compréhension des informations médicales pour un public plus large.

En résumé, ces solutions offrent des approches complémentaires pour simplifier les textes médicaux en anglais. Elles facilitent l'accès à l'information médicale en fournissant des résumés clairs et concis, ainsi qu'une génération de texte médical simplifié. De plus, l'analyse sémantique est prise en compte pour garantir la précision et la compréhension des termes médicaux complexes.

L'un des défis majeurs auxquels nous avons été confrontés réside dans les performances limitées des machines utilisés pour effectuer le fine-tuning nécessaire. Par ailleurs, la recherche d'un ensemble de données dans le domaine médical afin de mener des recherches et d'entreprendre des travaux a également constitué un obstacle majeur.

Nos perspectives futures se concentrent sur:

- La validation de nos approches en utilisant divers ensembles de données médicaux provenant de différentes langues.
- La normalisation des abréviations médicales, en établissant des règles claires et unifiées pour la conversion des abréviations en leur forme longue correspondante.
- L'intégration de ressources lexico-sémantiques telles que VerbNet et FrameNet offre une perspective prometteuse pour améliorer la qualité et la précision de la simplification des textes médicaux.

L'objectif est de permettre une diffusion à grande échelle, touchant un public diversifié à travers le monde, afin de favoriser une meilleure compréhension de la simplification des textes médicaux.

# Références


[1] Ornella Wandji Tchami. Analyse contrastive des verbes dans des corpus médicaux et création d'une ressource verbale de simplification de textes. 2018. Linguistique. Université de Lille; Universität Hildesheim, 2018. Français. ffNNT : 2018LILUH015ff. fftel 01998026f.

[2] Jehle D. V. Janicke D. M. Moscati. Lerner, E. B. Medical communication : do our patients understand ? *The American journal of emergency medicine, 18(7), 764–766*, 2000.

[3] Perceval Wajsbürt. Extraction and normalization of simple and structured entities in medical documents. *Santé publique et épidémiologie. Sorbonne Université,English. ⟨NNT : 2021SORUS541⟩.*, 2021.

[4] South B.R. Christensen L. et al. Mowery, D.L. Normalizing acronyms and abbreviations to aid patient understanding of clinical texts. *ShARe/CLEF eHealth Challenge 2013, Task 2. J Biomed Semant 7, 43*, 2016.

[5] Laurent Audibert. Outils d'exploration de corpus et désambiguïsation lexicale automatique. *Université de Provence - Aix-Marseille I, Français. ffNNT : ff. fftel-00004475f*, 2003.

[6] Gholamreza Haffari. Islam Nassar, Michelle Ananda-Rajah. Neural versus non-neural text simplification: A case study. *Faculty of Information Technology, Monash University, VIC, Australia. Department of Infectious Diseases, The Alfred Hospital and Central Clinical School.*, 2019.

[7] Clement Chatelain et Romain Herault. Découvrez le fonctionnement des réseaux de neurones récurrents., 2021. OpenClassrooms.

[8] Pierre Sémard. La transcription automatique dans le domaine médical[online], Aug. 2021. 94200 Ivry-sur-Seine France.

[9] Adrien Pupier. Cécile Macaire, Lucía Ormaechea. Une chaîne de traitements pour la simplification automatique de la parole et sa traduction automatique vers des pictogrammes. *29e conférence sur le Traitement Automatique des Langues Naturelles,Avignon, France. pp.111-123.*, Jun 2022.

[10] Jorge Gracia, Elena Montiel-Ponsoda, Philipp Cimiano, Asunción Gómez-Pérez, Paul Buitelaar, and John McCrae. Challenges for the multilingual web of data. *Journal of Web Semantics*, 11:63–71, 2012.

[11] Daniel Jurafsky and James Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. 2, 02 2008.

[12] Youssif Zaghwani and Youssif Omar. Effects of morphological levels on understanding meaning of words in english. *Conference: The 6th International Conference on Control Signal Processing*, 06 2019.

[13] William O'Grady and Videa P. De Guzman. Morphology: the analysis of word structure. *Contemporary linguistics, ed. by William O'Grady*, 2010.

[14] Joakim Nivre. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553, 2008.

[15] Chris Manning and Hinrich Schütze. Foundations of statistical natural language processing., May 1999.

[16] Daniel Jurafsky and James H. Martin. Speech and language processing. *3rd Draft Edition, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, January 7, 2023.

[17] Roberto Navigli and Simone Paolo Ponzetto. Building a very large multilingual semantic network. *Association for Computational Linguistics*, 2010.

[18] Xu Q Li F Rao G Tao C. Zhang L, Hu J. A semantic relationship mining method among disorders, genes, and drugs from different biomedical datasets. *BMC Med Inform Decis Mak. 2020;20(Suppl 4):283. Published 2020 Dec 14.*

[19] Jelena Jovanović and Ebrahim Bagheri. Semantic annotation in biomedicine: the current landscape. *Jovanović and Bagheri Journal of Biomedical Semantics (2017).*

[20] Hindawi. International journal of clinical practice. *Volume 2022, Article ID 6807484, 27 pages.*

[21] J. L. Mey. Pragmatics: An introduction. *Oxford: Blackwell.*, 2001.

[22] James F Allen and George Ferguson. Actions and events in interval temporal logic. *Journal of logic and computation*, pages 531–579, 1994.

[23] Sergei Nirenburg and Victor Raskin. Ontological semantics. 2004.

[24] Denis Carcagno Corinne Fournier. Karine Baschung, Gabriel G. Bès. Text generation. *[Research Report] Université Blaise-Pascal, Clermont-Ferrand; Alcatel Alsthom Recherche; Dassault Aviation.*, 1991.

[25] Shashi Narayan. Generating and simplifying sentences. computation and language [cs.cl]. *Université de Lorraine,⟨NNT : 2014LORR0166⟩. English.*, 2014.

[26] Ilya Sutskever, James Martens, and Geoffrey E Hinton. Generating text with recurrent neural networks. *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1017–1024, 2011.

[27] Zhifeng Chen Yonghui Wu, Mike Schuster and al. Google's neural machine translation system: Bridging the gap between human and machine translation, 2016.

[28] K. Fokou. Nlp  modèles de langage — smals research,, Juin 2019.

[29] Vincent Gardeux. Conception d'heuristiques d'optimisation pour les problèmes de grande dimension : application à l'analyse de données de puces à adn. *C. Autre [cs.OH]. Université Paris-Est. Français. ffNNT : 2011PEST1022ff.*, 2011.

[30] P. Alain. Evaluation des modèles de langage n-gram et n/m-multigram. *ACL Anthology.*, 2005.

[31] Lee Kenton Devlin Jacob, Chang Ming-Wei and Toutanova Kristina. BERT: Pre-training of deep bidirectional transformers for language understanding. *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)*, 2019.

[32] Shashi Narayan. Text summary evaluation based on interpretable semantic textual similarity. *Multimedia Tools and Applications, 2022, ⟨10.1007/s11042-022-14082-6⟩. ⟨hal-03839544⟩*, 2014.

[33] R. Bois. Introduction au résumé automatique,. *Le Data Blog, Oct. 19, 2021.*

[34] Mohamed Hedi Maaloul. Approche hybride pour le résumé automatique de textes. *Application à la langue arabe. Traitement du texte et du document. Université de Provence - Aix-Marseille I,*, 2012.

[35] Iti Mathur. Vaishali Gupta, Nisheeth Joshi. Design  development of a rule based urdu lemmatizer. *Department of Computer Science, Apaji Institute, Banasthali University, Rajasthan, India.*, 2016.

[36] Bastien Rance Xavier Tannier Aurélie Névéol. Nesrine Bannour, Perceval Wajsbürt. Privacy-preserving mimic models for clinical named entity recognition in french. *Journal of Biomedical Informatics, 2022, 130, pp.104073. ⟨10.1016/j.jbi.2022.104073⟩. ⟨hal-03655039⟩*, 2022.

[37] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics (Oxford, England)*, 36, 09 2019.

[38] Navya Kollapally and James Geller. Clinical biobert hyperparameter optimization using genetic algorithm clinical biobert hyperparameter optimization using genetic algorithm. 02 2023.

[39] Yingce Xia Tao Qin Sheng Zhang Hoifung Poon Renqian Luo, Liai Sun and Tie-Yan Liu. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *arXiv:2210.10341v3 [cs.CL] 3 Apr*, 2023.

[40] M. Karhade MD PhD. Page by page research review: Biogpt: Generative pre-trained transformer for biomedical text generation and mining,. *Medium, Mar. 02, 2023. [Online].*

[41] Yingce Xia and Renqian Luo. Biogpt: Generative language models for healthcare and beyond biogpt: Generative language models for healthcare and beyond - nlp summit,. *NLP Summit, Apr. 11, 2023.*

[42] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, November 2022.

[43] S. Lee A. Abbas and M. Afzal. Meaningful information extraction from unstructured clinical documents. *ResearchGate,*, Oct. 2019.

[44] Maxime E. Leroux D. Zrenner E. Sergouniotis, P. I. An ontological foundation for ocular phenotypes and rare eye diseases. researchgate. *University of Manchester and Manchester Royal Eye Hospital, Oxford Road,Manchester M13 9WL, UK,and for the ERN-EYE Ontology Study Group.*, 2019.

[45] Matthew Shardlow. A survey of automated text simplification. *Text Mining Group, School of Computer Science University of Manchester, Manchester, United Kingdom.* Email: mshardlow@cs.man.ac.uk.

[46] David Oniani et al. Toward improving health literacy in patient education materials with neural machine translation models. *Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA.*

[47] Lapata M. Zhang X. Sentence simplification with deep reinforcement learning [internet]. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017.*

[48] Gurevych I. Zhu Z, Bernhard D. A monolingual tree-based translation model for sentence simplification. *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). p. 1353–1361.*

[49] Remi Cardon. Approche lexicale de la simplification automatique de textes médicaux (lexical approach for the automatic simplification of medical texts). *In Actes de la Conférence TALN. Volume 2 - Démonstrations, articles des Rencontres Jeunes Chercheurs, ateliers DeFT, pages 159–174, Rennes, France. ATALA.*, 2018.

[50] Rémi Cardon. Simplification automatique de textes techniques et spécialisés. *Linguistique. Université de Lille, 2021. Français. ⟨NNT : 2021LILUH007⟩.*

[51] Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online, June 2021. Association for Computational Linguistics.

[52] Amal Menzli. Tokenization in nlp: Types, challenges, examples, tools. 2023.

[53] Jean-Pierre Jarry. Le traitement de texte et ses applications pédagogiques et didactiques. *Bulletin de l'EPI (Enseignement Public et Informatique), 1989, 54, pp.131-146,⟨edutice-00001248⟩.*, 2019.

[54] Mauro Di Pietro. Text summarization with nlp: Textrank vs seq2seq vs bart. *Article on Towards Data Science.*, March 2020.

[55] Stéphane Dufau Gala Núria Jacques Ginestié et al. Ludivine Javourey-Drevet, Núria Gala. Text simplification to improve reading fluency and comprehension. *Colloque SFERE- Provence : Apprentissage et Education Conditions, contextes et innovations pour la réussite scolaire, universitaire et professionnelle, Apr 2018, MARSEILLE, France. ⟨hal-01890345⟩.*, 2018.

[56] Maâli Mnasri. Résumé automatique multi-document et dynamique. *École doctorale n°580 Sciences et Technologies de l'Information et de la Communication (STIC) Spécialité de doctorat: Informatique*, 2018.

[57] Catherine Berrut. Une méthode d'indexation fondée sur l'analyse sémantique de documents spécialisés : le prototype rime et son application à un corpus médical. modélisation et simulation. 1988. Université Joseph-Fourier - Grenoble I. Français. ⟨NNT : ⟩. ⟨tel-00330027⟩.

[58] Le Beux Pierre. Pouliquen Bruno, Delamarre Denis. Indexation de textes médicaux par extraction de concepts,et ses utilisations. *European Commission, IPSC, Joint Research Centre – 21020 ISPRA - Italie2Laboratoire d'informatique Médicale – Faculté de Médecine - 35033 Rennes cedex – France.*, 2020.

[59] Radja Messai. Ontologies et services aux patients : Application à la reformulation des requêtes. *Informatique et langage [cs.CL]. Université Joseph-Fourier - Grenoble I, 2009. Français. ffNNT : ff. fftel-00952564f*, 2009.

[60] Aerin. Perplexity intuition (and its derivation)- towards data science. medium., (2022, September 19).

[61] F. Chiusano. Two minutes nlp, learn the rouge metric by examples. *Article on Medium.*, (2022, April 13).

[62] Ajitesh Kumar. Accuracy, precision, recall f1-score, python examples. Posted in Data Science, Machine Learning, Python, September 2022.