Etude de faisabilité de l'indexation automatique d'une base de connaissance

Côme Yarhi

le cnam
intd

Etude de faisabilité de l'indexation automatique d'une base de connaissance

Mémoire

pour l'obtention du Titre professionnel

« Chef de projet en ingénierie documentaire

et gestion des connaissances »

Niveau 7 – Bac+5

Date et lieu de la soutenance

- Paris. 15 décembre 2022

Membres du jury

- Loïc Lebigre, INTD
- Juliette Zwegers, Novartis

Promotion 52 (2021-2022)

YARHI Côme. Etude de faisabilité de l'indexation automatique d'une base de connaissance**.** Mémoire professionnel INTD, Titre 7, Chef de projet en ingénierie documentaire. Conservatoire national des arts et métiers – Institut national des Sciences et Techniques de la Documentation, 2022, 113p. Promotion 52.

Ce mémoire est une étude de faisabilité d'indexation automatique d'une base de connaissance d'un service de documentation chez un grand groupe pharmaceutique. Il fait un état des lieux sur les solutions d'intelligence artificielle pour traiter ce problème. Avec un focus sur le TAL et plus particulièrement le TAL biomédical. Il discute de l'état de l'art des différent modèles en mettant l'accent sur ceux ayant une architecture transformer. Les cas d'études présents dans le mémoire se base sur la littérature indexée dans PubMed. Le mémoire aborde des tâches comme celle de la reconnaissance d'entités nommées et celle de l'extraction de relations. Il traite de la problématique du bilinguisme. Enfin ils discutent plusieurs solutions possibles dont une faite maison.

Descripteurs
- Intelligence artificielle
- Apprentissage profond
- TAL
- TALN
- Plongement lexical
- PubMed
- Base de connaissance
- KM
- Métadonnées
- Indexation
- Automatisation
- Modélisation
- Propriété intellectuelle
- Biomédical

This master thesis is a feasibility study on how to realize the automation of indexing a knowledge base at a documentation service in a large pharmaceutical group. It makes an inventory of artificial intelligence solutions to deal with this problem. With a focus on NLP and more particularly biomedical NLP. This study discusses the state of the art of the different models with an emphasis on those with a Transfomer architecture. The case studies presented in this work are based on the literature indexed in PubMed. The study addresses tasks such as named entity recognition and relationship extraction. It deals with the issue of bilingualism. Finally, there is discussed several possible solutions including a homemade one.

Keywords
- Artificial Intelligence
- Deep Learning
- NLP
- Word embbedings,
- PubMed
- Knowledge Base
- KM
- Metadata
- Indexation
- Automatization
- Design
- Copyrights
- Biomédical

Je dédie ce mémoire à Jérémy Howard qui a, tel Prométhée, rendu au commun des mortels l'intelligence artificielle.

Table des matières

LISTE DES TABLEAUX

LISTE DES FIGURES

REMERCIEMENTS

Je remercie, chaleureusement, Juliette Zwegers et Mathieu Chansard pour leur accueil chez Novartis, leur bienveillance et leur patience à toute épreuve. Je remercie aussi les collègues de l'ICM et de toute la DAP que, j'ai pu rencontrer, pour m'avoir consacré un temps précieux. Un grand merci à Loïc Lebigre pour m'avoir accompagné tout au long de ce mémoire et pour m'avoir transmis le goût de la modélisation ! Un grand merci également à Nadia Raïs pour son dévouement et pour avoir débloqué des situations difficiles : merci d'être ce si merveilleux « couteau suisse » ! Merci aussi à Ghislaine Chartron et à Mathieu pour leur éclairages savants sur des points sourcilleux du droit européen et français. Je n'oublie pas mes grands-parents et ma mère pour leur soutien, une attention toute spéciale pour ma sœur : pour son affection et son écoute. Je remercie mon père pour sa façon toute à lui de me remettre devant mes responsabilités. *Last but not least*, mon fils, pour m'obliger à faire des pauses…En partageant avec lui mon ordinateur… pour qu'il puisse jouer à Fortnite…

INTRODUCTION

Pour le journal The Economist la nouvelle génération de l'intelligence Artificielle est la nouvelle "Frontière", il propose un éclairage grand public qu'il m'a paru intéressant de traduire et synthétiser ici. [1]

Les modèles de base (*foundation models or pretrained models*) sont la dernière version de « l'apprentissage profond (*deep learning* ou DL) » une technique qui a pris de l'importance il y a dix ans et maintenant domine le champ de l'intelligence artificielle (IA). Librement basés sur la structure en réseau des neurones du cerveau humain, les systèmes DL sont « entraînés » à l'aide de millions ou de milliards d'exemples de textes, d'images ou de clips audio. Ces dernières années, l'inflation du coût, en temps et en argent, de la formation de systèmes DL toujours plus grands avait fait craindre que la technique n'atteigne ses limites. Certains se sont inquiétés d'un « hiver de l'IA ». Mais les modèles de base montrent que la construction de DL toujours plus grand et plus complexe continue en effet à libérer de nouvelles capacités toujours plus impressionnantes. Personne ne sait où est la limite.

Les nouveaux modèles peuvent être réaffectés d'un type de problème à un autre avec une relative facilitée au moyen d'un réglage fin (*fine tuning*). C'est en raison de cette faculté, dans l'industrie, qu'ils sont souvent appelés « modèles de base ». Cette capacité à baser une gamme d'outils différents sur un seul modèle est une évolution non seulement de ce que l'IA peut faire, mais aussi de comment l'IA fonctionne en tant que business. « Auparavant, les modèles d'IA étaient très spéculatifs et artisanaux, mais maintenant leur développement est prévisible », explique Jack Clark, cofondateur d'Anthropic, une startup d'IA et auteur d'une newsletter très lue. « L'IA est en train d'entrer dans son ère industrielle. »

Plus proche de nos métiers de la documentation, la revue Archimag a elle aussi consacré un dossier à la veille et à l'intelligence artificielle qui justifie notre légitimité à entreprendre une telle étude de faisabilité. Voici l'introduction de leur dossier :

« À écouter les éditeurs de solutions de veille, l'intelligence artificielle dispose de nombreux atouts à faire valoir : elle permet de hiérarchiser l'information en détectant les mots les plus importants et elle est capable d'établir des relations entre des termes dispersés dans un texte.

Elle s'applique également à l'automatisation de filtres qui permettent de réduire le temps que les veilleurs consacrent à cette tâche à faible valeur ajoutée, y compris pour traiter des corpus textuels dans d'autres langues que le français (anglais, espagnol, arabe, portugais…).

Autres avantages : l'intelligence artificielle est en mesure de procéder à de la reconnaissance d'images et à de la recommandation de contenus. On pourrait ajouter la capacité de l'IA à thématiser des flux d'information en les regroupant par secteurs économiques, par pays ou par marchés.

Quasiment toutes les étapes du cycle de la veille peuvent tirer profit de l'IA. La collecte avec l'identification de nouvelles sources, la validation des sources selon de multiples critères (référencement, avis et commentaires, nombre d'articles publiés, construction du site…),

[1] The Economist.AI'S NEW FRONTIER, 11-17 juin 2022.p.11.

l'analyse (création de résumés sur la base des informations les plus pertinentes), la diffusion avec l'ajustement des livrables de veille selon les besoins de l'utilisateur final. »[2]

Enfin j'ai jugé intéressant d'inscrire cette étude dans une stratégie plus large de *knowledge Management*.

Les leviers de la transformation numérique dans le cadre du KM pour le développement d'un service ou d'une entreprise ont été listés dans un ouvrage rédigé par un spécialiste du KM qui fait référence en France. En voici le substrat :

« Définir la gestion des connaissances en entreprise ou *knowledge management* (KM) comme une association de la capitalisation et de la collaboration n'est plus suffisant. Dans le sillage de la transformation numérique des organisations, le KM doit plus s'intégrer dans les processus d'entreprise et se présenter comme un levier de transformation. En effet il faut reconnaître que le KM bénéficie des démarches de transformation numérique. Car la transformation numérique, c'est fondamentalement un changement de comportement et une montée en compétence dans l'usage de l'information.

Le KM gère des informations. Il questionne donc les professionnels de l'information et souvent les bouscule. Par ses solutions techniques (portail KM, bases de connaissances,), il s'inscrit dans un paysage que les démarches de gestion globale des contenus visent à concevoir et à administrer. Un KM opérationnel inscrit de fait ses outils dans le système d'information des organisations, ses bases dans la cartographie de l'ECM (Enterprise Content Management).

Parce qu'il est devenu l'un des processus des organisations, parce qu'il ne fonctionne pas tout seul, parce qu'il a besoin d'être piloté par des enjeux métiers et managé localement, le KM a besoin de nouveaux métiers.

Il intègre une transformation de ses acteurs dans leurs comportements de partage des connaissances, mais aussi de relecture de leurs activités pour y trouver du sens (démarche de réflexivité). Pour faire du KM, il faut donc être autant psychologue qu'ingénieur dans son approche. »[3]

Les dernières années ont vu une augmentation rapide du nombre d'articles scientifiques dans le domaine biomédical. Ces documents sont pour la plupart disponibles et facilement accessibles sous forme électronique. La connaissance du domaine qui y est incorporé est essentielle pour la recherche et les applications biomédicales, ce qui rend les techniques d'exploration et d'exploitation de cette littérature biomédicale très exigeantes. De nombreux efforts ont été déployés sur ce sujet par les communautés du biomédical et *du Machine Learning*. La communauté du biomédical se concentre davantage sur les problèmes d'application concrets et préfère donc des méthodes plus interprétables et descriptives, tandis que la communauté du *Machine Learning* recherche davantage des performances supérieures et une capacité de généralisation. C'est cette double dynamique qui conduit au développement

[2] https://www.archimag.com/veille-documentation/2021/05/07/intelligence-artificielle-veille-augmentee Consulté le 22/08/2022

[3] Chastenet de Géry G. Le knowledge management : Un levier de transformation à intégrer. De Boeck Supérieur, 2018.p.13-14

de modèles de plus en plus sophistiqués et universels, ou très spécifique au domaine biomédical[4].

L'objectif de cette étude est de fournir un examen des avancées récentes en matière de *machine learning* dans le domaine du biomédical d'inspirer de nouvelles orientations de recherche, pour résoudre le problème concret d'automatisation de l'indexation de la base de connaissance d'IVAN, à l'ICM chez Novartis.

En effet, la problématique de mon mémoire est de réaliser une étude de faisabilité d'indexation automatique ou semi-automatique d'une base de connaissance. Car, le processus actuel d'indexation peut largement être rationnalisé : il nécessite beaucoup de saisie et il est redondant. L'indexation compte pour 20% de l'activité des documentalistes et cumulé à l'activité de veille, en amont, à 40% de l'activité totale.

Le manager et les documentalistes scientifiques souhaitent allouer leur temps là où elles ont le plus de valeur ajoutée. L'idée est d'intégrer des outils de veille comme Curebot® en amont. Et, en aval, identifier des solutions d'intelligence artificielle déjà sur le marché ou de faire développer une solution, *ad hoc*, sur mesure, en interne, dans le but d'automatiser l'indexation. Pour cela il est nécessaire de bien qualifier le problème à traiter et de documenter les caractéristiques et les spécifications techniques nécessaires à sa résolution.

Dans une première partie : après une présentation générale de Novartis et des activités de l'ICM, j'analyserai l'existant : Les outils, les *workflows*, les processus mis en œuvre pour indexer les articles dans la base de connaissance : IVAN, par les documentalistes scientifiques, c'est-à-dire la nature de leur travail et les pistes déjà explorées pour une indexation semi-automatique.

Dans un second temps je proposerai une qualification de la nature du problème et des éléments, à la fois techniques et méthodologiques, permettant sa résolution. Je n'oublierai pas de mettre en avant les forces et les faiblesses d'un tel projet.

1 CADRE DE TRAVAIL ET ETAT DES LIEUX DE L'ACTIVITE DES DOCUMENTALISTES SCIENTIFIQUES DE L'ICM

1.1 NOVARTIS : UNE ENTREPRISE A ENTITE GLOBALE

Novartis est une firme multinationale qui se promeut comme une entreprise pharmaceutique innovante et à forte assise scientifique qui « réinvente la médecine pour améliorer et prolonger la vie des gens » et se targue d'être à la pointe pour résoudre certains des problèmes de santé les plus sévères de la société en développant des traitements « révolutionnaires » tout en réfléchissant à comment les rendre disponibles au plus grand

[4] Zhao S., Su C., Lu Z., *et al*. Recent advances in biomedical literature mining. Brief Bioinform, 2021

nombre de patients. Novartis traite la plupart des principaux domaines pathologiques : du cancer aux maladies cardiaques en passant par les maladies génétiques rares. Ses médicaments ont été consommés par 766 millions de patients et vendus dans environ 155 pays, à travers le monde, en 2021.[5]

Le laboratoire Novartis, a deux divisions opérationnelles mondiales : *Innovative Medicines*, spécialisée dans les médicaments protégés par des brevets, et Sandoz, qui commercialise des génériques et des biosimilaires. Ces divisions s'appuient sur des équipes de recherche et développement, des opérations de fabrication, des services commerciaux, des technologies et des fonctions support d'entreprise.

Depuis cet été, une nouvelle organisation se met en place- Côté Monde : il existe désormais l'entité *Innovative Medicines US* d'une part, et l'entité *Innovative Medicines international* d'autre part, dont l'Europe fait partie. Voici ci-dessous, l'organigramme du nouveau CoDir[6] :

[5] Novartis. Annual Report, 2021

[6] https://www.novartis.com/sites/novartis_com/files/novartis-org-chart.pdf. Consulté le 15/11/2022

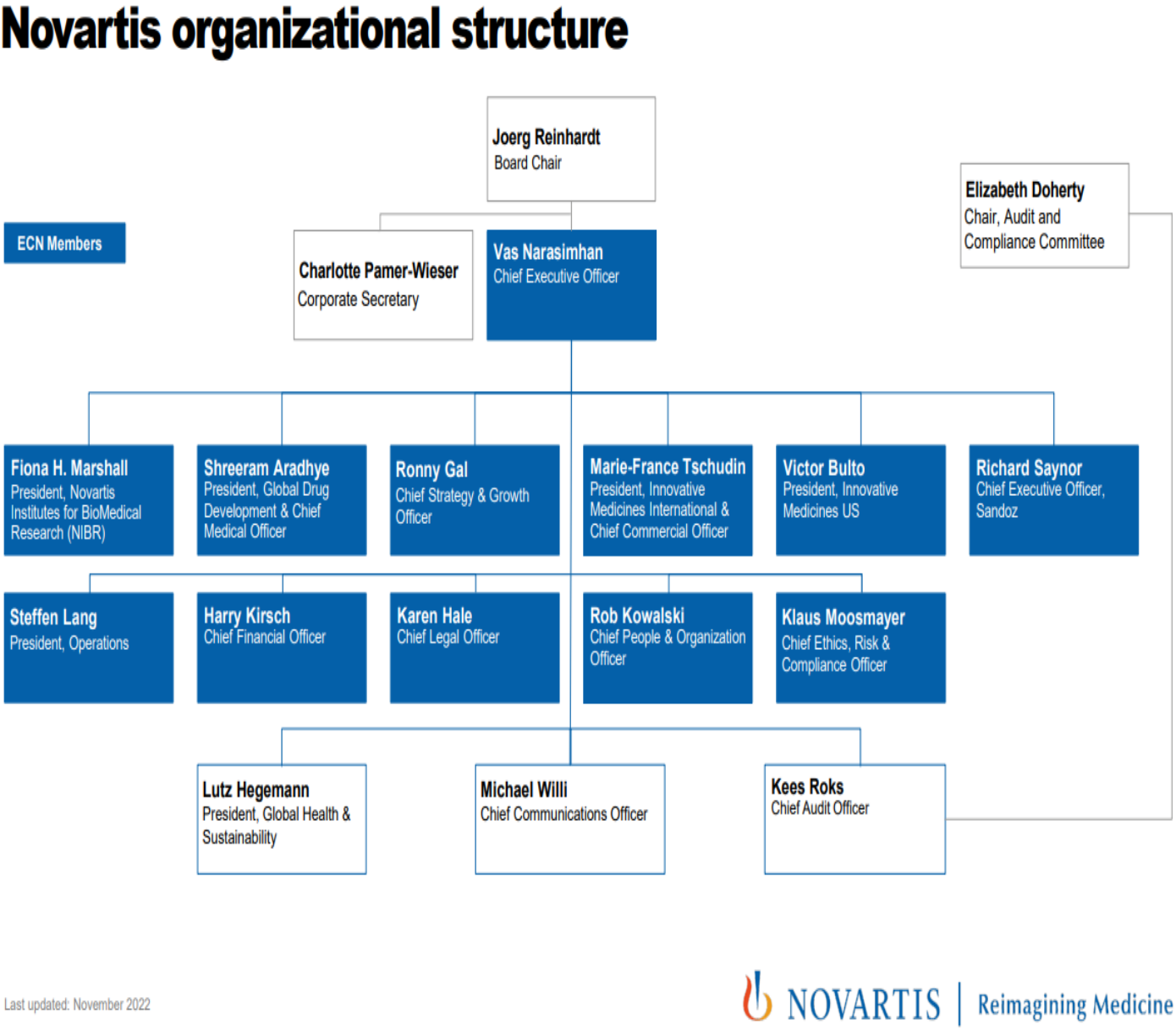


Figure 1 : Novartis CoDir

Innovative Medicines Us et *International* proposent des traitements dans le champs des maladies cardiovasculaires, de l'hématologie, des tumeurs solides, de l'immunologie, des neurosciences, de l'ophtalmologie et des troubles respiratoires. Ces deux organisations sont à la pointe des thérapies génique pour les maladies, rares et sévères, génétiques et neurologiques[7].

[7] https://www.novartis.com/about/innovative-medicines. Consulté le 15/11/2022

Figure 2 : La structure du groupe Novartis

Le dynamisme de l'entreprise, l'épicentre de la création de valeur, est drainé par la recherche et développement (R&D)[8] : les instituts Novartis pour la recherche biomédicale (*the Novartis Institutes for BioMedical Research* ou NIBR) sont les moteurs de l'innovation. Ces innovations sont mises en production par le *Global Drug Development* (GDD) qui supervise le développement des médicaments découverts par les chercheurs et les partenaires de Novartis et le NTO (*Novartis Technical Operations*) qui fabriquent les produits et les dispositifs pour que Novartis et Sandoz les livrent, à leurs clients, à travers le monde. Cela est fluidifié par le département *Customer & Technology* Service (CTS) qui s'occupe de toutes les activités du *Digital* et ses outils qui standardisent et rendent performants les process métiers. Enfin toutes

[8]Voir le graphique, ci-dessus, mis à jour par, mes soins, à partir du graphique original du rapport annuel de Novartis, 2021

les fonctions, *Corporate,* vitales pour un grand groupe comme Novartis sont dans des domaines d'expertises que sont la finance, les ressources humaines, le juridique, la communication ou bien la santé mondiale et l'éthique, ou les risques et la conformité.

Dans le monde, Novartis emploie 108 514 personnes (104 323 postes équivalents temps plein), avec environ un cinquième de ses employés travaillant dans la recherche et le développement[9].

Le siège social de Novartis est situé à Bâle, en Suisse. Il existe plus de 380 sites dans le monde entier.[10]

Les principaux sites de Novartis par nombres d'employés et aires géographiques d'importance sont[11] :

Figure 3 : Les principaux sites de Novartis dans le monde

1.1.1 NOVARTIS EN FRANCE

1.1.1.1 LES PRINCIPAUX SITES EN FRANCE[12]

[9] Novartis. Annual Report, 2021

[10] Ibid.

[11] Ibid.

[12] Carte, ci-dessous issue de la présentation par les ressources humaines de Novartis France, lors de la journée d'intégration.

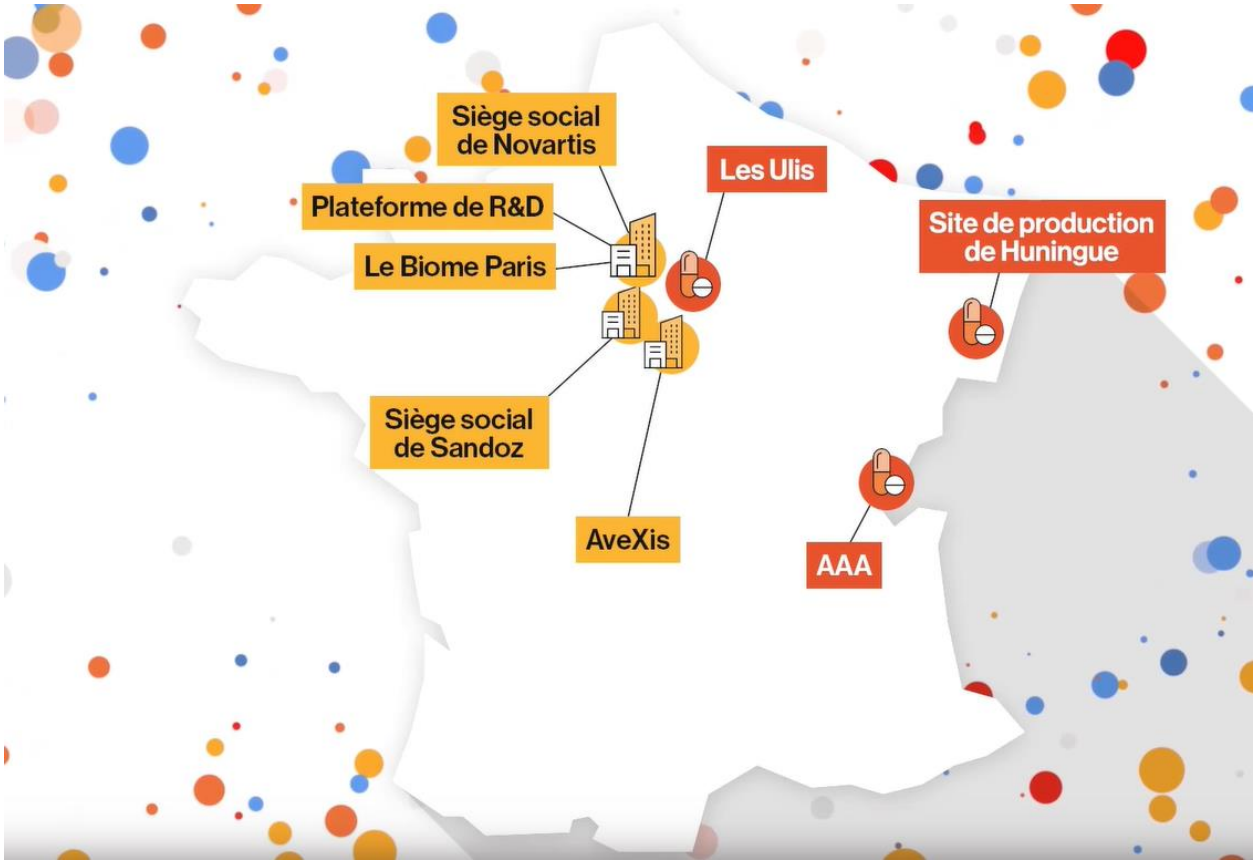


Figure 4 : Les principaux sites de Novartis en France

1.1.1.2 QUELQUES CHIFFRES POUR PRESENTER L'ENTREPRISE[13].

[13] https://www.novartis.com/fr-fr/novartis-en-france Consulté le 17/11/2022

Figure 5 : Novartis un acteur majeur en France

1.1.1.3 LIEU DU STAGE

Nous nous situons au siège social de Novartis France à Rueil-Malmaison (92) où la Direction des Affaires Pharmaceutiques (DAP) à laquelle je suis rattaché est basée. La DAP gère toutes les opérations pharmaceutiques de Novartis France, de la fabrication au contrôle de la promotion. Elle est dirigée par le Pharmacien Responsable assisté de trois Directrices qui l'accompagnent dans ses missions quotidiennes. Je vais présenter rapidement l'organisation et les principales fonctions de cette direction essentielle à Novartis France.

Tout d'abord à la tête de la DAP le Pharmacien Responsable. Il organise et supervise l'ensemble des opérations pharmaceutiques de Novartis Pharma SAS France. C'est un poste clé, puisqu'il partage la responsabilité légale avec le Directeur Général de Novartis France et engage même sa responsabilité pénale personnelle. Ses missions principales :

- Fabrication, dont la libération des lots
- Suivi, dont la gestion des réclamations, les rappels éventuels, etc.
- Stockage
- Distribution, dont les conditions de transport
- Importation et exportation
- Pharmacovigilance
- Information médicale
- Promotion : publicité et formation de la visite médicale

Ensuite, si on descend d'un échelon dans l'organigramme on trouve les trois Directions des Affaires Pharmaceutiques.

La Direction de la Qualité de l'Information et des pratiques assure la bonne conformité des informations distribuées aux professionnels de santé. Ses activités recouvrent :

- Le bon usage du médicament (BUM) :
  - Validation des éléments de communication aux professionnels de santé et des supports de formation pour les visiteurs médicaux.
  - Élaboration de ces éléments avec les équipes Médical & Marketing.
- L'activité d'information promotionnelle :
  - Organisation des activités liées à la promotion du portefeuille Novartis, afin de maintenir la certification de la force de vente.
- Le Service Bonnes Pratiques :
  - Vérification de la conformité des opérations Novartis (siège et terrain) avec la loi d'encadrement des avantages (LEA) et de leur soumission pour approbation auprès des autorités compétentes.
  - Formation de l'ensemble des collaboratrices et collaborateurs.

On trouve également à ce même niveau la Direction Assurance Qualité. Au quotidien, ses équipes interagissent avec les autres entités du Groupe et les différents services ayant un impact sur les activités pharmaceutiques. Elles assurent notamment la qualité :

- Des produits commercialisés et des médicaments expérimentaux :
  - Libération des lots
  - Suivi des réclamations
  - Interactions avec les autorités de santé
  - Interface avec le dépositaire de Novartis
- Des processus de l'entreprise en relation avec
  - Les Bonnes Pratiques de Fabrication,
  - Distribution,
  - Clinique,
  - Pharmacovigilance :
    - Coordination des audits et des inspections des autorités de santé,
    - Gestion des procédures,
    - Suivi des sous-traitants du périmètre de responsabilité pharmaceutique.

Enfin, y figure, la direction de l'Information et Communication Médicale (ICM), le service où se passe mon stage. Cette Direction gère et diffuse l'information médicale en réponse aux demandes des Professionnels de Santé et des patients.

- Les Experts en Information Médicale :

- Cette équipe traite 15 000 demandes par an, avec le maximum de personnalisation et de réactivité. Un coordinateur en charge de la gestion du Call Center et du CRM (*Customers Relation Managemen*t) optimise l'efficacité du service.

➢ Les Documentalistes Scientifiques :

- Ce service traite 1000 demandes par an, sélectionne et diffuse près de 1 900 articles et dépêches scientifiques par an, fournissant ainsi un support de connaissances et d'analyse de l'information aux Professionnels de Santé ainsi qu'aux collaborateurs Novartis. C'est dans ce service que j'effectue mon stage.

1.2 PRESENTATION DES MISSIONS DE L'ICM

L'ICM a une responsabilité pharmaceutique pour les produits *NOVARTIS Innovative Medicines* commercialisés ou en développement et sur leur environnement. Ses missions sont :

➢ Répondre aux demandes d'informations des clients externes (Professionnels de Santé, Patients) 24h/24 et 7j/7 et rappeler le bon usage du médicament

➢ Sélectionner, analyser et diffuser les informations scientifiques pertinentes et actualisées en interne

Les métiers de l'ICM sont :

➢ Coordinateur Call Center

- Assure l'interface entre Novartis et le prestataire externe en charge du call center :
  - Administrateur de l'outil CRM
  - Partage des informations en temps réel sur leurs produits.
- Contrôle qualité des demandes enregistrées dans l'outil CRM
- Veille au respect du cahier des charges.
- *Reporting* des activités de l'ICM

➢ Assistante spécialisée

- Administrateur de la boite générique ICM
  - Transmet les demandes aux services concernés,
  - Assure la réponse aux Professionnels de santé,
  - Envoi des brochures
- Abonnements aux revues :
  - Suivi des contrats,
  - Réclamations,
  - Traitement des factures.
- Gestion et actualisation de la liste des produits Pharma commercialisés.

- Activités administratives et backup coordinateur Call Center.
- Exports des réconciliations bimensuelles Pharmacovigilance / Qualité produit.

➢ Experts Information et Communication Médicales
- Des missions auprès des professionnels de santé et des patients :

    - Réponses médicales argumentées orales ou écrites
    - Données cliniques, interactions, hors AMM, alternatives, populations spéciales, pharmacocinétique/pharmacodynamique (PK/PD) etc.
    - Rédaction des Réponses/Lettres Types pour les questions récurrentes.

- Une implication transversale

    - (Direction des Affaires Réglementaires, DAP, Pharmacovigilance (PV), Marketing (MKT), Médical, Logistique, Commercial...)

➢ Documentalistes Scientifiques et Chargée d'Administration en Documentation
- Des missions, également auprès des Professionnels de santé et des patients

    - Répondre aux demandes d'information avec délais et modes de réponse adapté
    - Recherches bibliographiques,
    - PGRs et brochures non promotionnelles.
    - Respect des bonnes pratiques de documentation, des droits de copie et des règles de la Transparence.

- Des missions auprès des collaborateurs
    - Répondre aux demandes PV, Market Access (réinscription), et BUM (réunions Suivi des prescriptions et utilisations non conformes des médicaments).
    - Réaliser et diffuser une veille bibliographique sur leurs produits, pathologies associées et concurrents (mailing, Alertes IVAN « L'lnfo Comme on I'M »).
    - Gérer le fonds documentaire ICM (abonnements aux revues électroniques/papier, matériel éducationnel dans le cadre des Plans de Gestion des Risques (PGR) et brochures non promotionnelles).
    - Former à la recherche bibliographique, aux sources d'information, aux droits de copie.

1.3 CHAINE DOCUMENTAIRE DE L'ICM

1.3.1 ANALYSE DE L'EXISTANT

1.3.1.1 UN EXEMPLE DE PROCESSUS TYPE DE VEILLE ET D'ANALYSE EN VUE D'UNE INDEXATION DANS LA BASE DE CONNAISSANCE, À PARTIR D'UN REX[14] D'UNE DOCUMENTALISTE SCIENTIFIQUE :

<u>Consignes :</u>

Pouvez-vous me faire une synthèse personnelle de la manière dont vous travaillez pour sélectionner les articles pertinents qui vous intéressent pour les indexer dans IVAN ? ***Je propose un document collaboratif et évolutif Word que je partage pour centraliser, éviter les répétions et donc de vous faire perdre votre temps.***

- **Quelles informations vous extrayez, vous ciblez ?**
    - ***Vos trucs et astuces pour filtrer l'info, repérer et déterminer ce qui a de l'intérêt***

- **Essayez de conscientiser un maximum la manière dont vous procédez et de l'expliciter par écrit.**
    - ***Par exemple :***
        - ***À ma connaissance, deux documentalistes commencent par faire un ctrl-F pour voir si la DCI (la molécule), la pathologie de leur franchise est citée***
        - ***Une des documentaliste regarde les graphes et les légendes en priorité. Pourquoi ?***

- **Vous regardez tous si le produit est mature, les effets indésirables et l'hors AMM. Et le reste ?**

[14] Retour sur expérience

- **Quels sont les termes, les mots-clés, phrases/symptômes/traitements/problématiques/dispositifs/patients ciblés etc…**

- **Enfin, si vous aviez à nouveau le temps de faire une veille des concurrents, qu'est-ce que vous surveilleriez ?**

Chaque détail, geste, intuition, étape compte. Essayez d'être le plus complet possible même avec des détails qui pourraient paraître insignifiants comme le "ctrl-F" cité en exemple qui ont, en fait, une grande importance.

1) **Quelles informations vous extrayez, vous ciblez ?**

La documentaliste interviewée : Nous avons connaissance des pathologies à surveiller, à repérer, ainsi que les centres d'intérêt des collaborateurs internes. Je peux classer les infos à collecter en deux types :

- Les articles "produit" : ils ont essentiellement pour sujet les indications de nos produits commercialisés, mais aussi les pathologies pour lesquelles il existe des études cliniques en cours pour une molécule en développement, issue de la recherche. Cela peut être aussi un effet indésirable rapporté dans l'article, ou encore un usage hors-indication du produit. Le nom de la molécule est alors cité dans l'article.
- Les articles "environnement" : exemple : données épidémiologiques France, Europe, qui sont rares et difficiles à trouver, alors je sélectionne. Cela peut être aussi des recommandations de stratégies thérapeutiques dans une pathologie suivie, de la part de sociétés savantes françaises. Cela peut être aussi le résultat d'enquête de vraie vie pour les patients traités, avec parcours de soins, observance au traitement...

Si je pars d'un sommaire de revue, je lis dans un premier temps les titres des articles, puis je jette un coup d'œil aux auteurs (français ? Expert dans son domaine ?), et enfin, le résumé/abstract de l'article quand il y en a un. Après, selon le sujet (indications, pathologies, environnement thérapeutique), je vais consulter le texte intégral de l'article.

2) **Essayez de conscientiser un maximum la manière dont vous procédez et de l'expliciter par écrit.**

La documentaliste interviewée : Je parcours le texte de la publication, en lisant les titres des paragraphes et en me concentrant davantage sur les parties "méthodes, thérapies, traitements". Je regarde systématiquement les tableaux et infographies, car, souvent c'est un bon récapitulatif des traitements étudiés, (avec produits cités), et des données contenues dans l'article. Je lis également les légendes des photos, si elles existent, car parfois elles contiennent des infos sur

les patients et leurs traitements. Dans tous les cas, je lis rapidement et de manière efficace (œil exercé).

Enfin, je fais en effet un CTRL-F pour vérifier si je ne suis pas passée à côté du nom d'une molécule recherchée. Mais le CTRL-F seul n'est pas suffisant, car le nom de la molécule peut être mal orthographié (anglais-français) ou encore tronqué, comme en fin de phrase, avec le retour à la ligne. (Exemple : se*cu-*

kinumab)

3) **Vous regardez tous si le produit est mature, les effets indésirables et l'hors AMM. Et le reste ?**

La documentaliste interviewée : oui. Nous regardons s'il y a des informations de tolérance, et de hors-AMM lié à un de nos produits.

4) **Enfin, si vous aviez à nouveau le temps de faire une veille des concurrents, qu'est-ce que vous surveilleriez ?**

La documentaliste interviewée : Les résultats des études cliniques publiés et les publications de type "revue de littérature", "review", "synthèse" écrites sur un médicament concurrent.

5) **Quels sont les termes, les mots-clés, phrases, les symptômes, les traitements, les problématiques, les dispositifs, les patients ciblés etc… ?**

Réponse par l'exemple sur un processus complet de lecture rapide et sélective qui démontre comment un œil exercé de documentaliste scientifique à l'ICM s'y prend pour identifier, sélectionner un article de sa franchise et le qualifier en vue d'extraire les métadonnées pour son indexation et donc sa classification.

Ci-dessous un article[15] annoté par La documentaliste interviewée et mis en forme par mes soins.

À noter qu'en temps normal, les documentalistes n'ont pas le temps d'annoter, elles font tout de tête, à l'œil.

[15] © Brenaut, E. Prise en charge du psoriasis chez le patient souffrant d'obésité. Réalités Thér Dermato-Vénérol, 2022, 311(Cahier 1):7-8+10-1

Prise en charge du psoriasis chez le patient souffrant d'obésité

Page : 1

RÉSUMÉ : La prévalence de l'obésité est en nette augmentation dans le monde depuis plusieurs années. Les patients atteints de psoriasis souffrent plus souvent d'obésité que la population générale. Ce lien entre psoriasis et obésité est probablement bidirectionnel. En effet, l'obésité amplifie l'inflammation systémique et constitue ainsi un facteur de risque indépendant de psoriasis. On sait aussi que les patients avec un psoriasis ont des habitudes de vie qui pourraient favoriser le surpoids (isolement social, régime alimentaire moins équilibré, activité physique réduite).
Le dépistage de l'obésité doit se faire régulièrement lors du diagnostic et du suivi, ainsi que le dépistage du syndrome métabolique.
Les traitements systémiques sont moins efficaces chez les patients souffrant d'obésité et les effets secondaires plus fréquents. Certains traitements, comme les anti-TNFα, sont associés à une prise de poids.

problématique exprimé sur le sujet de l'article

mot-clé

mot-clé

rubrique cochée dans la grille

important à mettre en mot-clé, lien avec le psoriasis

E. BRENAUT
Service de Dermatologie, CHU de BREST ;
Laboratoire Interactions Épithéliums Neurones,
EA4685, BREST

L'obésité, une comorbidité fréquente du psoriasis

L'obésité est une comorbidité fréquente du psoriasis. Elle se définit par un IMC supérieur à 30 et se divise en :
– obésité modérée (grade 1) : IMC 30,0-34,9 ;
– obésité sévère (grade 2) : IMC 35,0-39,9 ;
– obésité morbide (grade 3) : IMC ≥ 40.

On estime qu'entre 11 et 25 % des patients atteints de psoriasis sont obèses [1]. En France, une étude observationnelle multicentrique a inclus 2 210 patients avec un psoriasis (en

les patients psoriasiques était supérieure à la fréquence dans la population générale, indépendamment de l'âge et du sexe, et la différence s'accentuait après 45 ans [2].

Ce lien entre obésité et psoriasis est probablement bidirectionnel. En effet les patients psoriasiques présentent plus fréquemment que dans la population générale un isolement social, des symptômes dépressifs, un régime alimentaire moins équilibré et une activité physique réduite pouvant favoriser l'obésité. Cependant, des études épidé-

Page : 2

poids conséquente (supérieure à 15 kg) avaient plus de risques de développer un psoriasis : RR 1,88 (1,44-2,46) [3].

Une autre étude a suivi plus de 33000 individus pendant 10 ans. En comparaison avec les sujets ayant un poids normal, les sujets obèses avaient un risque de 1,87 (IC95 % : 1,38-2,52) de développer un psoriasis. Une prise de poids supérieure à 10 kg et un tour de taille élevé étaient aussi associés au psoriasis [4].

Physiopathologie : le lien entre obésité et psoriasis

L'obésité est un facteur de risque indépendant de l'apparition du psoriasis par le biais de l'inflammation systémique. L'adiposité semble être un facteur central dans cette association et l'expansion du tissu adipeux majore l'inflammation psoriasique par la surexpression de médiateurs pro-inflammatoires par les adipocytes : les adipokines. Il s'agit de cytokines classiques telles que le TNFα, et des molécules spécifiques telles que la leptine et la résistine. Des concentrations sériques d'IL17 et IL23 plus élevées ont été mises en évidence chez des patients obèses. À l'inverse, la sécrétion d'adiponectine, adipokine anti-inflammatoire, est diminuée [5]. Par ailleurs, la libération continue d'acides gras libres dans la circulation systémique provoque un

génétique pourrait être suspecté car le gène *HLA-Cw6*, gène majeur de susceptibilité au psoriasis, serait aussi associé à l'obésité [5].

Impact de l'obésité sur les traitements

L'obésité est un élément à prendre en compte pour le choix d'un traitement. Les effets indésirables des traitements sont plus fréquents chez les patients obèses. Cela s'explique notamment par l'association de l'obésité avec le syndrome métabolique ou la stéatose hépatique : augmentation de l'hépatotoxicité du méthotrexate, augmentation de la néphrotoxicité de la ciclosporine [7].

Dans les différentes études, la réponse aux traitements systémiques et biologiques était diminuée chez les patients obèses [7]. Une méta-analyse a évalué l'impact de l'obésité sur l'efficacité des anti-TNFα chez des patients avec différentes maladies inflammatoires. Les patients obèses avec un psoriasis ou un rhumatisme psoriasique avaient un risque augmenté de 57 % d'être en échec de traitement (OR 1,57 ; IC95 % : 1,30-1,89) [8]. Dans l'étude ECLIPSE, la réponse au traitement était diminuée chez les patients obèses. À la semaine 48, le PASI 90 était atteint par 88,1 % des patients avec un poids normal traités par guselkumab *versus* 75,2 % de

l'IMC sur les traitements systémiques conventionnels et biologiques [11]. En analyse multivariée, une augmentation de 5 unités de l'IMC était associée à un risque accru de 12 % d'arrêter le traitement en raison d'un manque d'efficacité et de 17 % d'avoir un effet indésirable, quelle que soit la molécule utilisée. Les données du registre BioCAPTURE permettaient d'analyser la persistance du traitement. Un IMC élevé était prédictif d'un arrêt de l'étanercept et l'ustekinumab pour manque d'efficacité [12]. Dans le registre PSOBIOTEQ, les prescriptions de biothérapies étaient différentes selon que le patient était obèse ou non (molécules les plus prescrites : adalimumab, ustekinumab et étanercept), la persistance du traitement était diminuée, principalement à cause d'un manque d'efficacité [13].

Impact des traitements du psoriasis sur le poids

Une méta-analyse a analysé l'effet des biothérapies sur le poids. Six études ont été incluses représentant 862 patients et évaluant les anti-TNFα, l'ustekinumab et le secukinumab [14]. L'adalimumab était associé à la plus grande prise de poids, suivi de l'infliximab puis de l'étanercept. L'ustekinumab et le secukinumab n'avaient pas d'effet sur le poids. Lors des essais cliniques

traitements = Therapeutique en rubrique

Paragraphe sur la physiopath' : -> rubrique

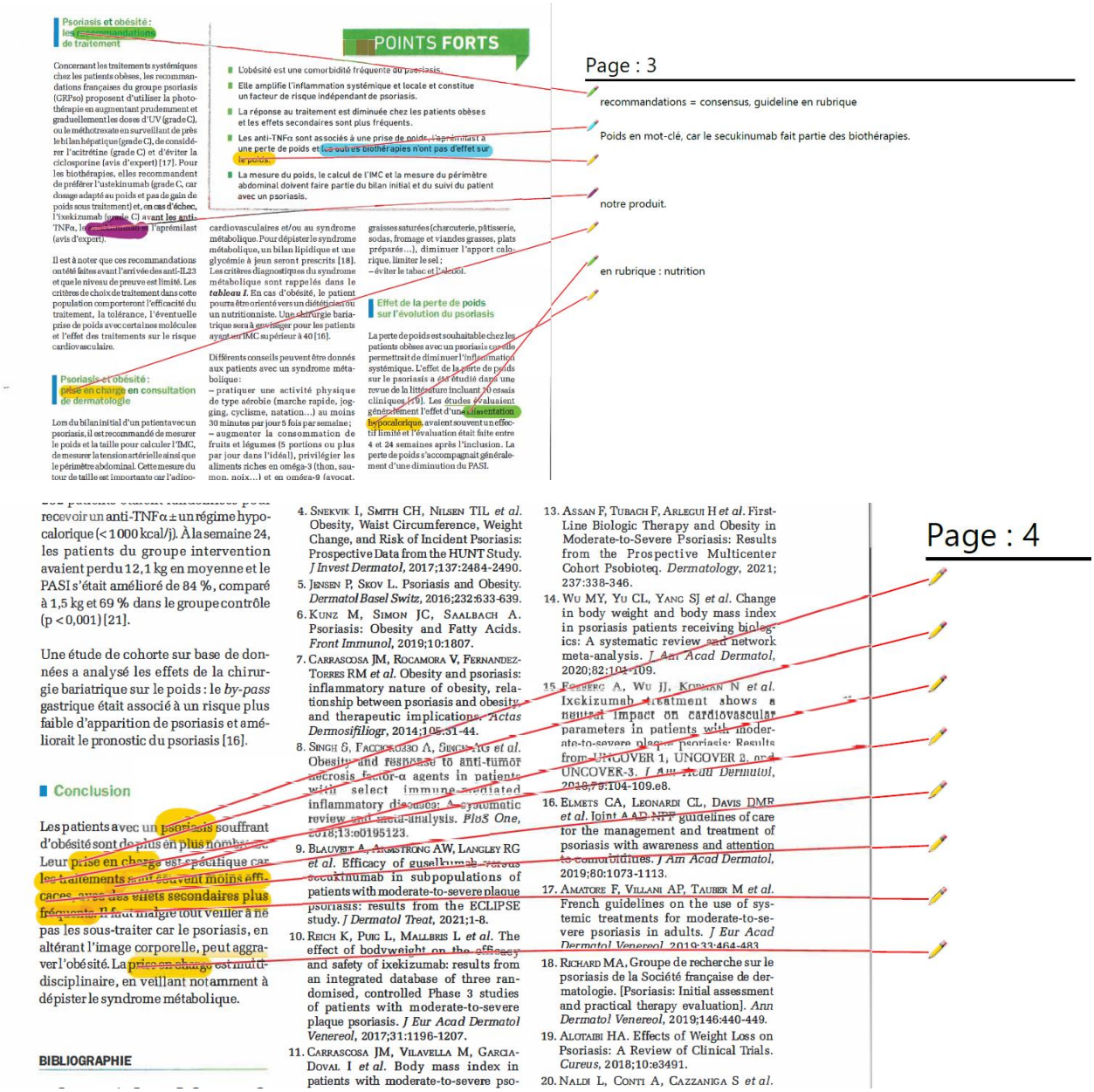


Figure 6 : Exemple d'un article annoté en vue d'une indexation dans IVAN

1.3.1.2 PRESENTATION DE LA BASE DE CONNAISSANCE DE L'ICM : IVAN ET DE SES USAGES

L'objectif d'une base de connaissance est tout simplement de rendre accessible des connaissances.

S'inscrivant dans une politique de *Knowledge Management* (ou gestion des connaissances), une base de connaissance constitue une source de documentation fiable et exhaustive sur un

sujet spécifique. Ce qui la différencie d'une base de données généraliste, susceptible de regrouper des informations variées.

Pour simplifier, une base de connaissance est comparable à une foire aux questions qui recensent les principales interrogations qu'un partenaire est susceptible de se poser sur un domaine donné. Toutefois, la base de connaissance ne survole pas un sujet. Elle l'approfondit afin de s'imposer comme une source exhaustive et de référence. Elle doit être en mesure d'apporter une réponse claire et intelligible face à une problématique donnée ou une explication initiale jugée trop évasive.[16]

1.3.1.2.1 *LES ILLUSTRATIONS QUI SUIVENT FONT UN TOUR COMPLET DE LA BASE, DE SES FONCTIONNALITES ET DE SES USAGES.*[17]

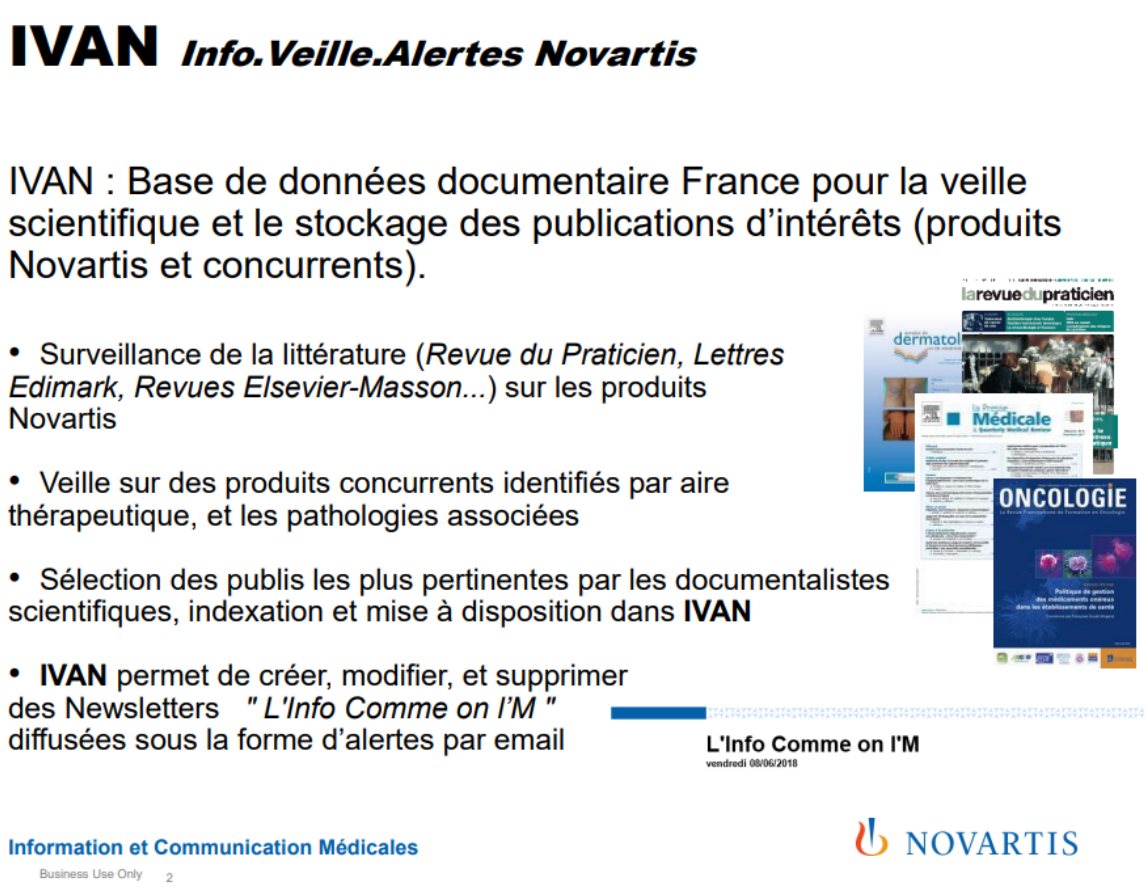


Figure 7 : À quoi sert IVAN ?

[16] https://www.journaldunet.fr/web-tech/guide-de-l-intelligence-artificielle/1501317-base-de-connaissance/ Consulté le 22/08/2022

[17] Elles sont issues d'une présentation montrée en interne à l'intention des collaborateurs de Novartis.

IVAN est hébergé sur le réseau Novartis.

- On peut y effectuer une recherche de publications sur une période donnée, une aire thérapeutique donnée, une revue précise... l'acronyme d'une étude clinique, ou tout simplement un article précis. Exemples de recherches :
  - « Je veux des articles français parus en 2015-2017 sur la BPCO chez les patients asthmatiques. »
  - « Je recherche cet article : Dossier : Dégénérescence maculaire liée à l'âge. Auteur : Bousquet. Revue du Praticien MG, 2018 ; 32(994):67-72 »

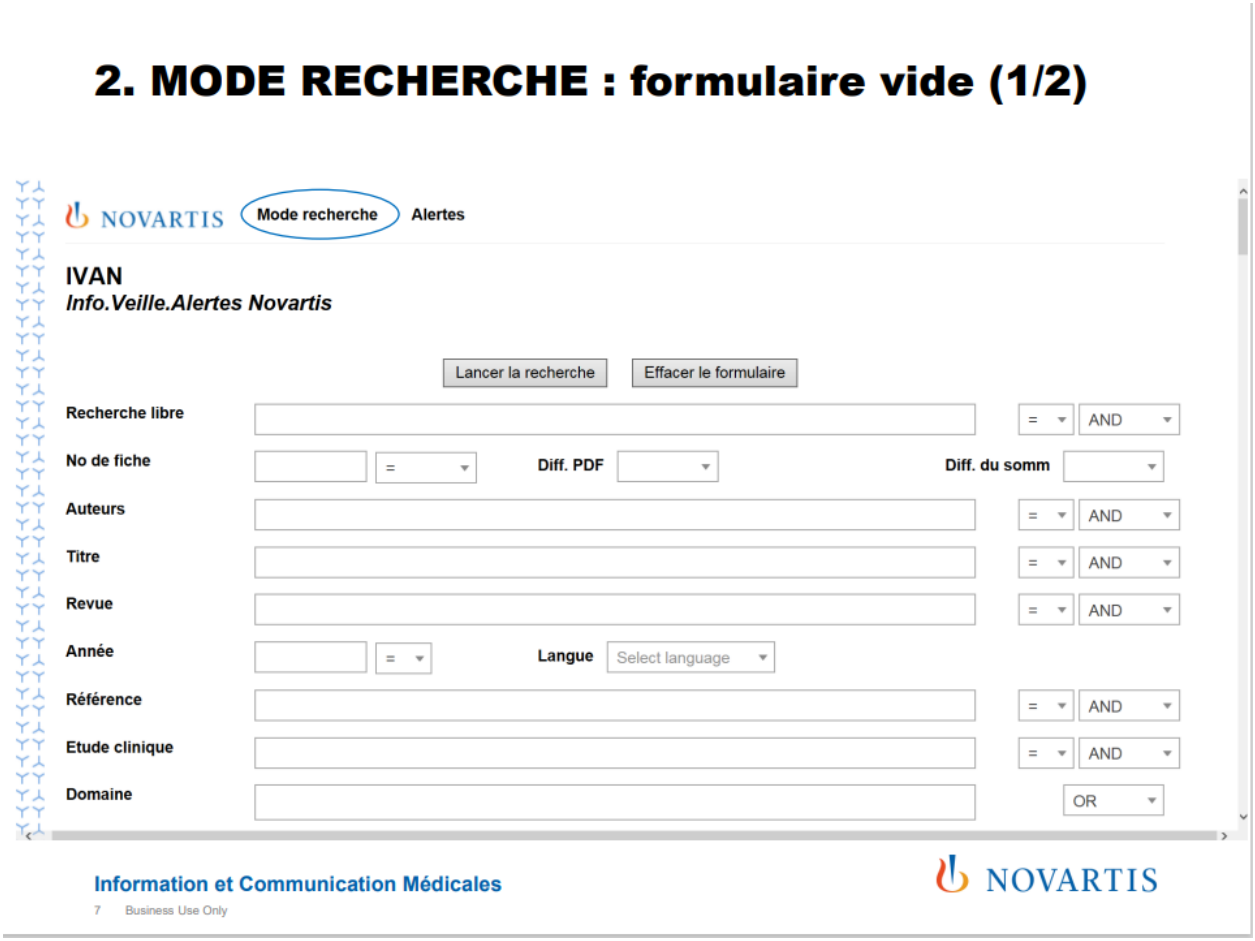


Figure 8 : Vue 1 du formulaire de recherche d'IVAN

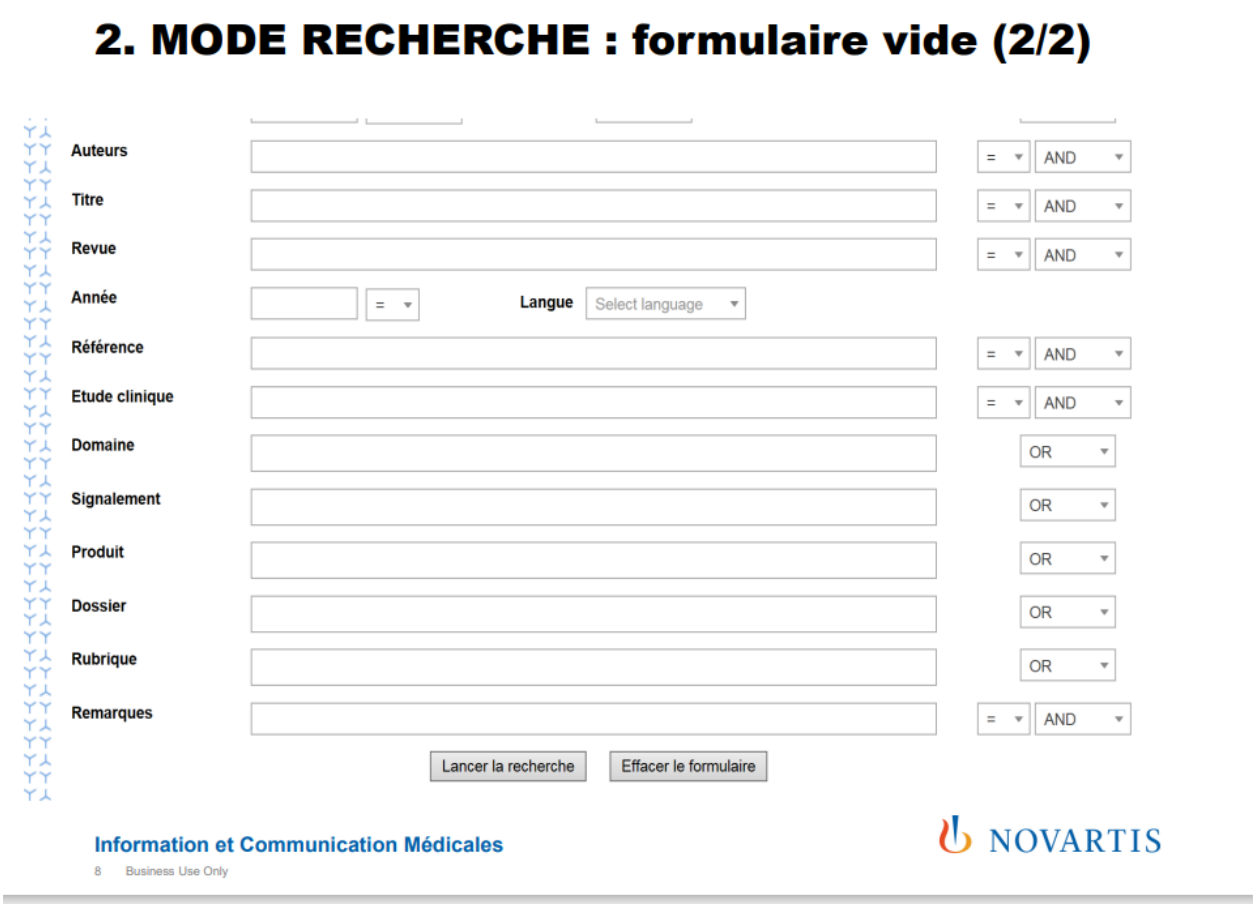


Figure 9 : Vue 2 du formulaire de recherche d'IVAN

- **MODE RECHERCHE—Deux différents modes de requêtes possibles :**
    1. La recherche via le champ « Recherche libre » s'effectue dans toutes les données associées à l'article. Exemples : Auteurs, Titre, Produit, texte intégral, etc.
    2. La recherche par mots-clés s'effectue via les champs menus déroulants : revue, dossier, rubrique, domaine, signalement, produit etc.
        - Recherche par mots-clés
            - Les mots-clés sont des données d'indexation intégrées par les documentalistes dans les champs : Domaine, Signalement, Produit, Dossier, Rubrique, Langue.
            - Sélection possible d'un ou plusieurs mots dans ces listes.
            - L'auto-complétion existe.

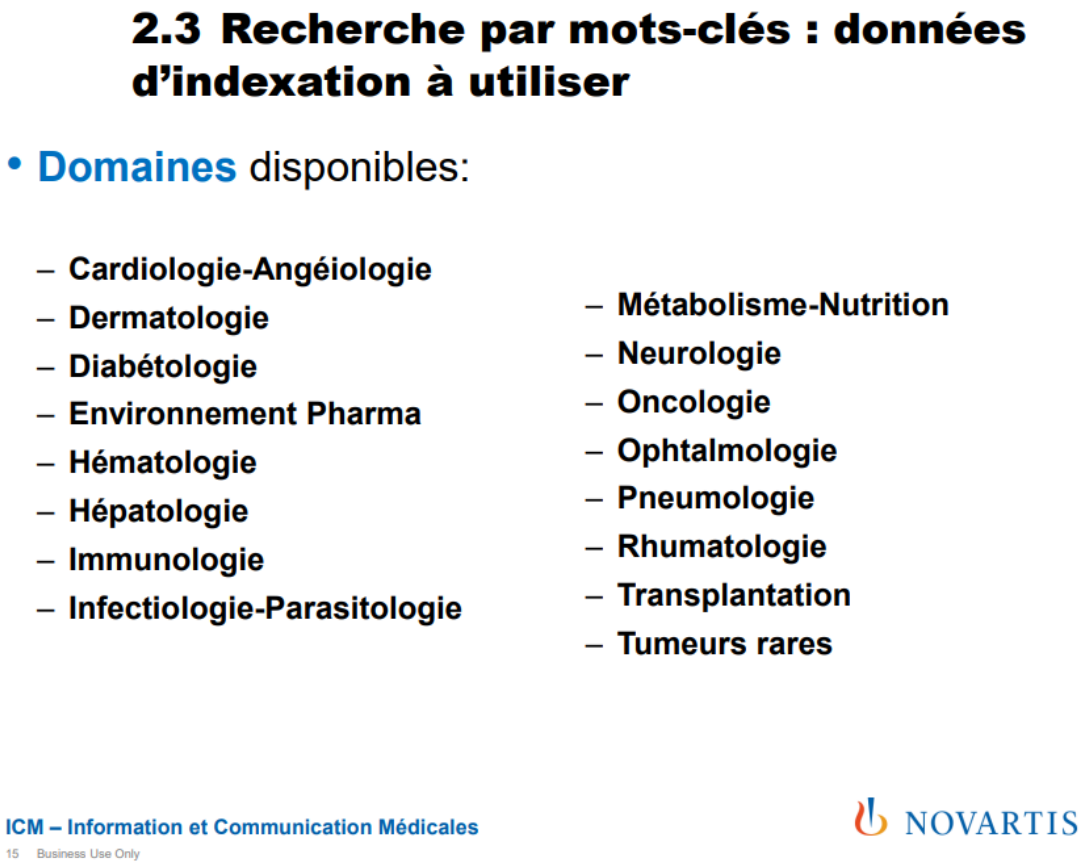


Figure 10 : Mots clés disponibles pour la recherche, ici les domaines

- **Recherche par mots-clés : Données d'indexation qu'on peut utiliser :**
  - Définition des Signalements :
    - Pharmacovigilance : toxicité, tolérance du produit ou de la classe, excepté la toxicité des concurrents.
    - Recommandations : guidelines, consensus, Haute Autorité de Santé (HAS) etc.
    - Revues générales : articles généraux sur les pathologies concernées par les produits Novartis, sur l'environnement des produits (Traité EMC, Expert Opinion, etc.)
    - Autres : articles sur l'environnement pharmaceutique (sur les décrets, essais cliniques etc.).

2.3 Recherche par mots-clés : données d'indexation à utiliser

- **Rubriques** disponibles :

Rubriques		
AMM		Mécanisme d'action
Biologie moléculaire		Médicaments Assoc. synergique
Chirurgie		Médicaments Comparaison
Chronobiologie		Médicaments Interaction
CI, Mises en garde, Précautions		Méthode de dosage
Complications		Méthodologie
Concurrence		Nutrition
Consensus/Guidelines		Observance
Diagnostic		Pharmacocinétique
Dispositifs		Pharmaco-économie
Echelles/Tests		Pharmacologie Clinique
Education du patient		Physiopathologie
Effets indésirables		Posologie, Mode d'emploi
Epidémiologie		Prévention
Essais cliniques		Qualité de vie
Essais cliniques (sous-groupe)		Réglementation
Examens biologique/paraclinique		Revue générale
Facteurs de risque		Sélectivité
Facteurs pronostiques		Sujet âgé
Fonction cognitive		Sujet jeune (pédiatrie)
Génériques		Sujet sportif
Génétique		Surdosage/Mésusage
Grossesse, allaitement		Switch
Hors AMM		Technique médico-chirurgicale
Imagerie		Thérapeutique
Insuffisant cardiaque		Thérapie ciblée
Insuffisant rénal		Tolérance
Interactions autres		Toxicologie

Information et Communication Médicales

17 Business Use Only

NOVARTIS

Figure 11: Mots clés disponibles pour la recherche, ici les domaines

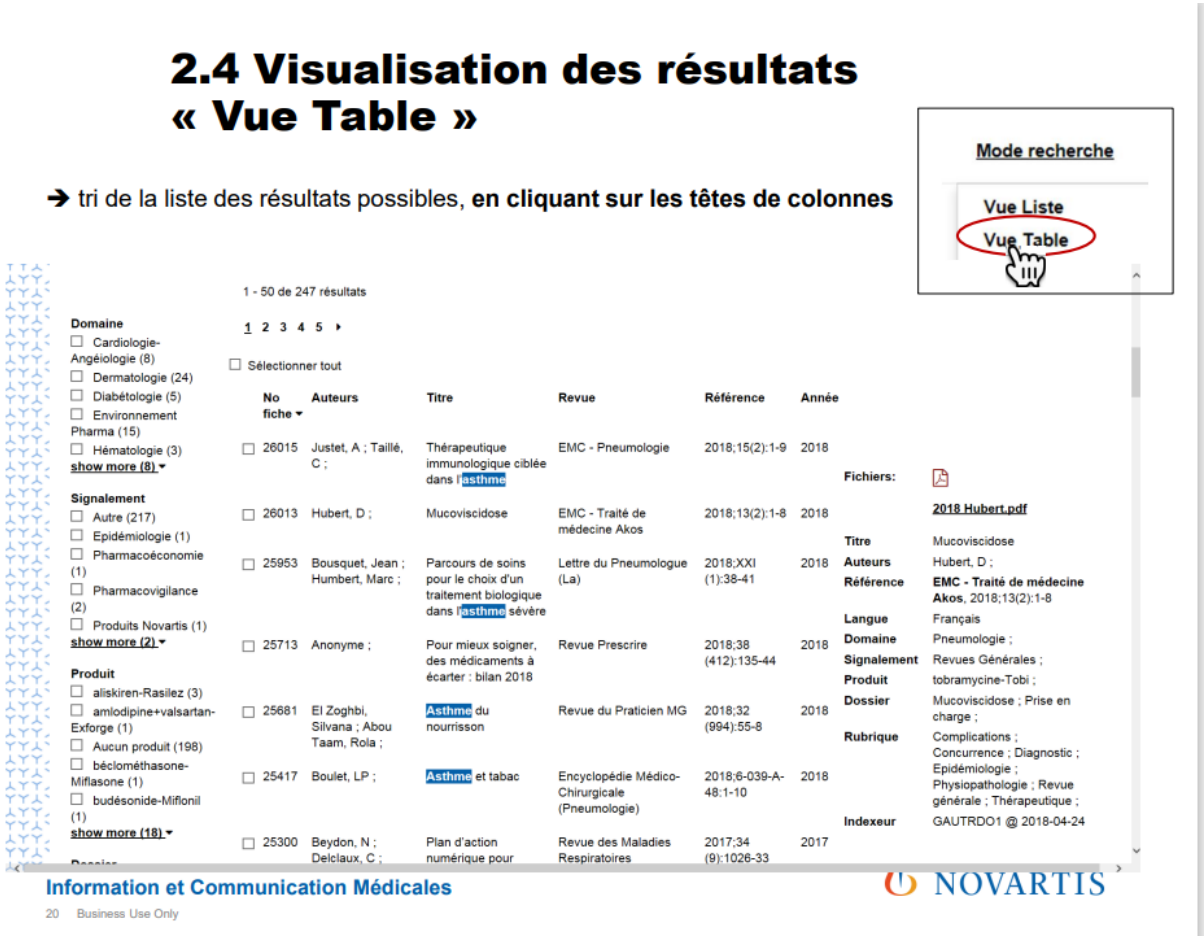


Figure 12: Visualisation des résultats : « vue table »

2.4 Visualisation des résultats : informations détaillées de l'article

Il faut survoler le titre de l'article avec le pointeur de la souris pour afficher à droite de la page les informations détaillées de cet article.

➡ Si les droits de copie spécifiques à cet article le permettent, le fichier PDF de l'article est disponible dans cette vue détaillée et accessible en cliquant sur le nom du fichier.

➡ Sinon, un lien intitulé **Demander l'article** permet de contacter l'ICM par email afin d'obtenir une copie de l'article.

➡ Le sommaire en pdf d'une revue peut aussi être visible ; le lien intitulé **Demander l'article du sommaire** permet de recevoir un article choisi de ce sommaire.

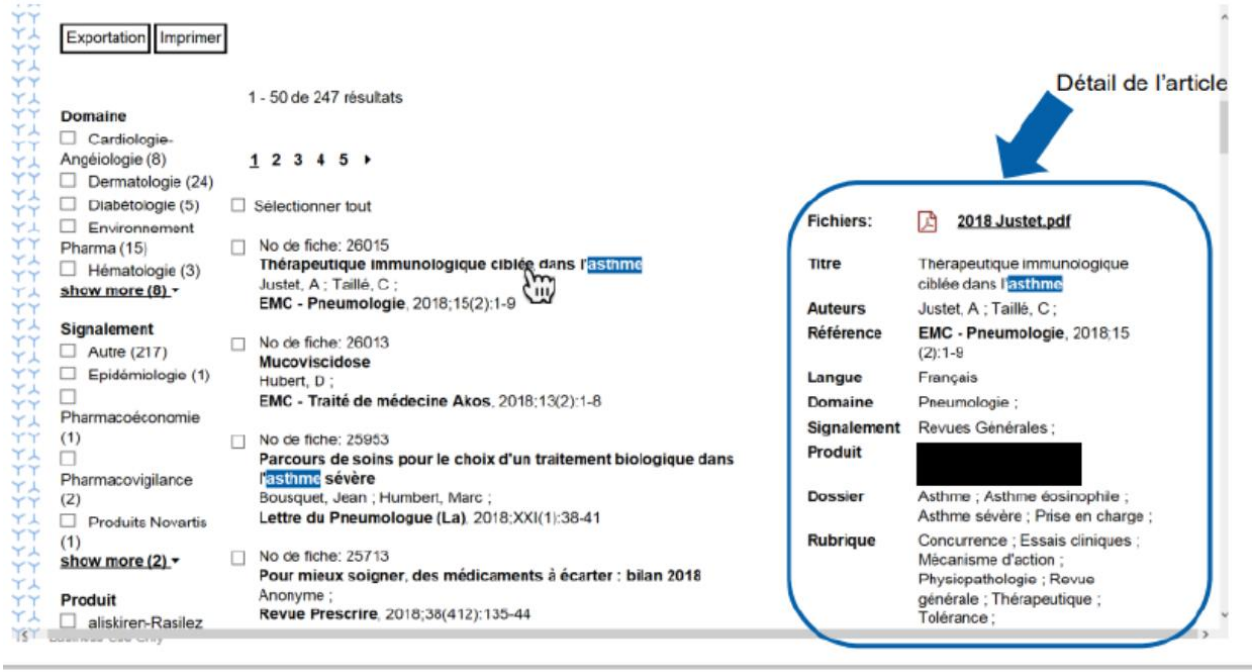


Figure 13: Visualisation des résultats : informations détaillées de l'article

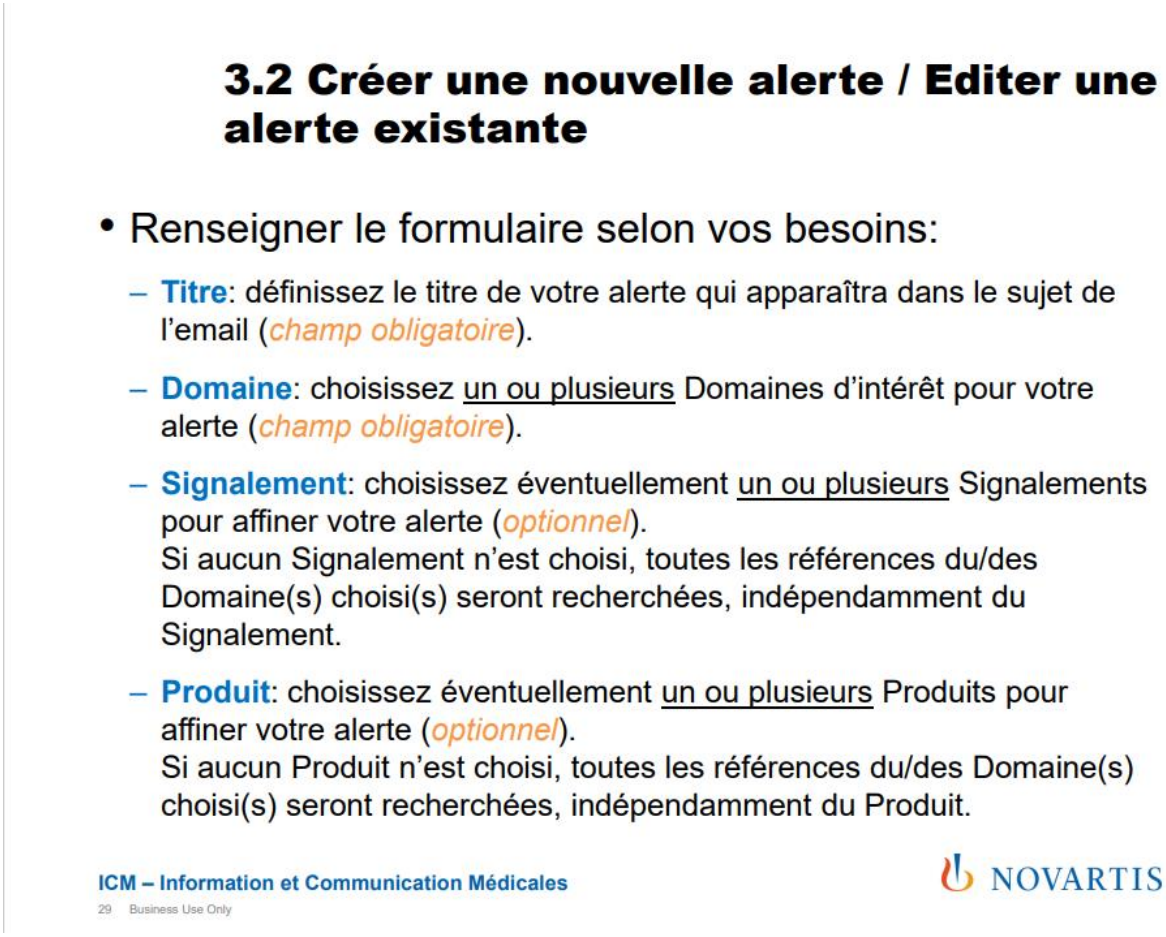


Figure 14 : Comment créer un alerte dans IVAN ?

- Fréquence et périodicité avec laquelle sont envoyés les alertes par emails :
  - Quotidien
  - Hebdomadaire
  - Bimensuel
  - "Mensuel

- Alertes paramétrables pour les collaborateurs internes à Novartis pour qu'ils puissent être informés des dernières saisies.

1.3.1.2.2 LA RAISON D'ETRE D'IVAN

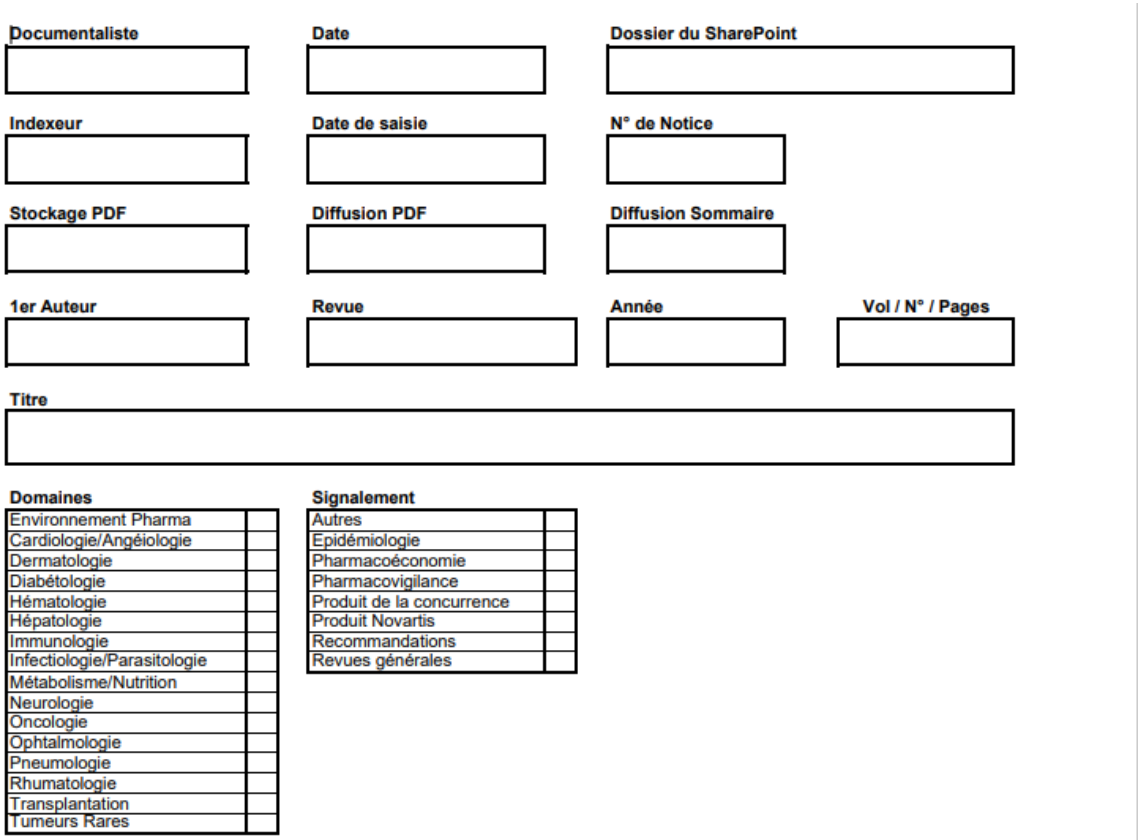
- Les alertes internes

- Stocker tout ce qui concerne la pharmacovigilance (PV) et le Hors Autorisation de Mise sur le Marché (AMM) dans le cadre du Suivi des prescriptions et utilisations non conformes des médicaments
    - De manière générale indexation et archivage pour garantir la traçabilité et rendre les recherches ultérieures possibles en cas de de demande des institutions de santé française telles que l'Agence nationale de sécurité du médicament (ANSM)

- Permettre le requêtage de tout ce qui a été publié dans la littérature française (au moins un auteur français ou en langue française).

1.3.1.2.3 LES GRILLES D'INDEXATION

C'est une indexation manuelle qui se fait sur un fichier Excel dont voici les vues :



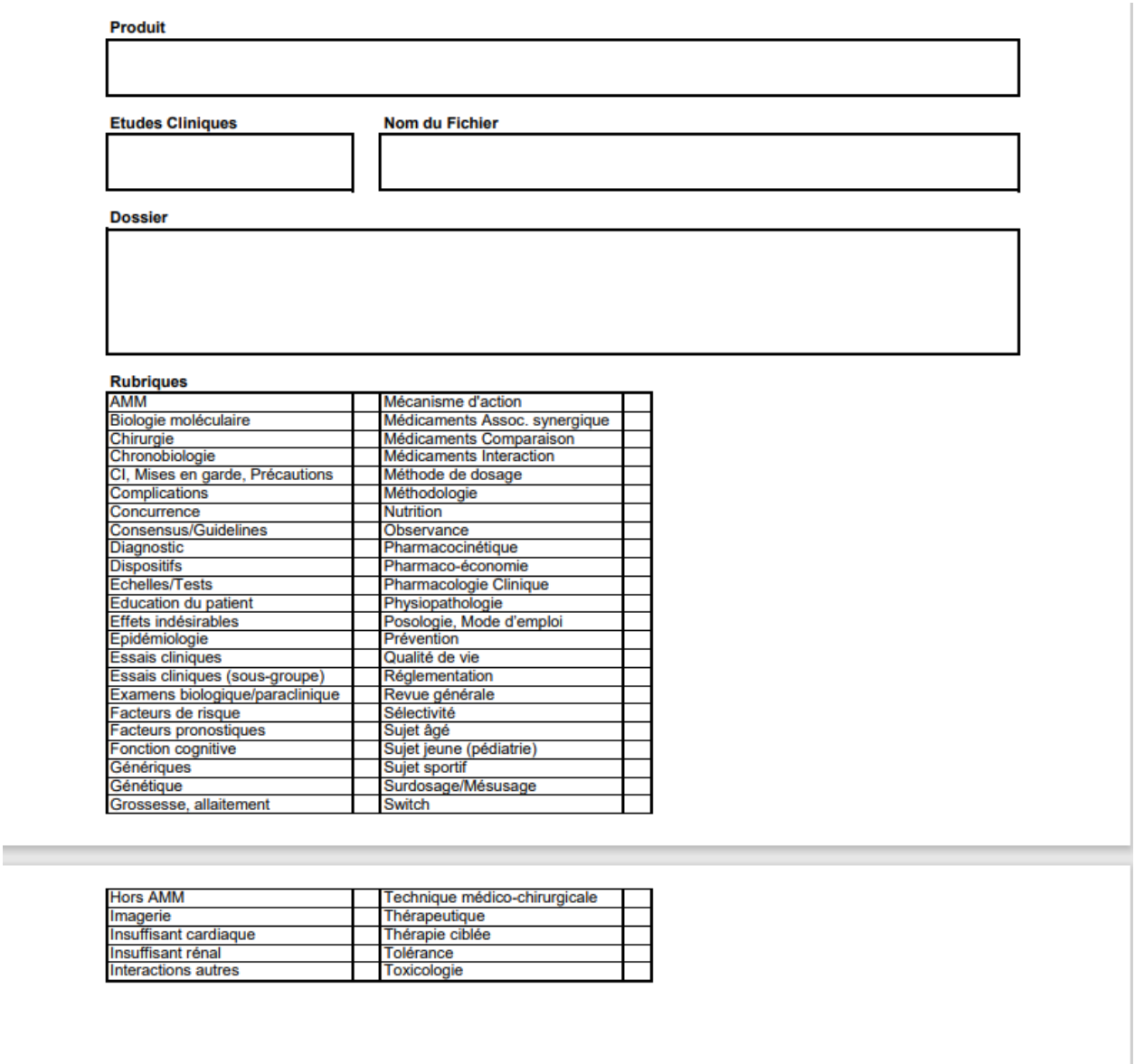
Documentaliste	Date	Dossier du SharePoint	
Indexeur	**Date de saisie**	**N° de Notice**	
Stockage PDF	**Diffusion PDF**	**Diffusion Sommaire**	
1er Auteur	**Revue**	**Année**	**Vol / N° / Pages**

Titre

Domaines	
Environnement Pharma	
Cardiologie/Angéiologie	
Dermatologie	
Diabétologie	
Hématologie	
Hépatologie	
Immunologie	
Infectiologie/Parasitologie	
Métabolisme/Nutrition	
Neurologie	
Oncologie	
Ophtalmologie	
Pneumologie	
Rhumatologie	
Transplantation	
Tumeurs Rares	

Signalement	
Autres	
Epidémiologie	
Pharmacoéconomie	
Pharmacovigilance	
Produit de la concurrence	
Produit Novartis	
Recommandations	
Revues générales	

Figure 15 : Grille d'indexation : vue 1



Produit

Etudes Cliniques

Nom du Fichier

Dossier

Rubriques

AMM		Mécanisme d'action	
Biologie moléculaire		Médicaments Assoc. synergique	
Chirurgie		Médicaments Comparaison	
Chronobiologie		Médicaments Interaction	
CI, Mises en garde, Précautions		Méthode de dosage	
Complications		Méthodologie	
Concurrence		Nutrition	
Consensus/Guidelines		Observance	
Diagnostic		Pharmacocinétique	
Dispositifs		Pharmaco-économie	
Echelles/Tests		Pharmacologie Clinique	
Education du patient		Physiopathologie	
Effets indésirables		Posologie, Mode d'emploi	
Epidémiologie		Prévention	
Essais cliniques		Qualité de vie	
Essais cliniques (sous-groupe)		Réglementation	
Examens biologique/paraclinique		Revue générale	
Facteurs de risque		Sélectivité	
Facteurs pronostiques		Sujet âgé	
Fonction cognitive		Sujet jeune (pédiatrie)	
Génériques		Sujet sportif	
Génétique		Surdosage/Mésusage	
Grossesse, allaitement		Switch	

Hors AMM		Technique médico-chirurgicale	
Imagerie		Thérapeutique	
Insuffisant cardiaque		Thérapie ciblée	
Insuffisant rénal		Tolérance	
Interactions autres		Toxicologie	

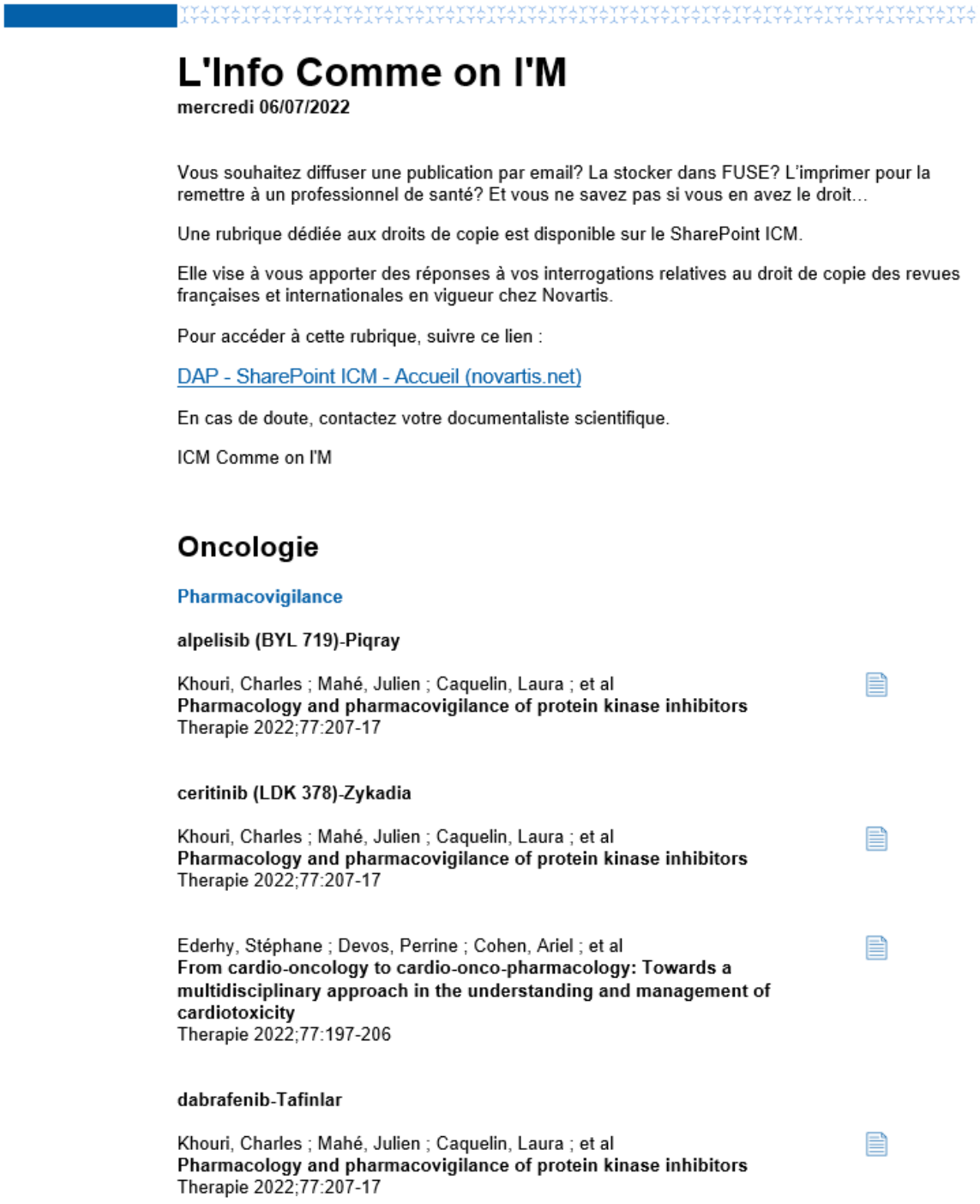
Figure 16: Grille d'indexation : vue 2

Les rubriques correspondent partiellement aux *subheadings* du MeSH.

1.3.1.2.4 DIFFUSION EN INTERNE

Les demandes d'articles, pour les articles qui ne sont pas en accès libre, sont traitées au cas par cas par l'ICM. Les autres sont diffusés au sein d'une newsletter qui s'appelle : **« L'Info**

comme on l'M ». Voici, ci-dessous, un exemple de newsletter reçue par un collaborateur qui a paramétré ses alertes selon les mots-clés et la fréquence choisis, et qui reçoit ainsi les derniers articles saisis dans IVAN.



L'Info Comme on l'M

mercredi 06/07/2022

Vous souhaitez diffuser une publication par email? La stocker dans FUSE? L'imprimer pour la remettre à un professionnel de santé? Et vous ne savez pas si vous en avez le droit…

Une rubrique dédiée aux droits de copie est disponible sur le SharePoint ICM.

Elle vise à vous apporter des réponses à vos interrogations relatives au droit de copie des revues françaises et internationales en vigueur chez Novartis.

Pour accéder à cette rubrique, suivre ce lien :

DAP - SharePoint ICM - Accueil (novartis.net)

En cas de doute, contactez votre documentaliste scientifique.

ICM Comme on l'M

Oncologie

Pharmacovigilance

alpelisib (BYL 719)-Piqray

Khouri, Charles ; Mahé, Julien ; Caquelin, Laura ; et al
Pharmacology and pharmacovigilance of protein kinase inhibitors
Therapie 2022;77:207-17

ceritinib (LDK 378)-Zykadia

Khouri, Charles ; Mahé, Julien ; Caquelin, Laura ; et al
Pharmacology and pharmacovigilance of protein kinase inhibitors
Therapie 2022;77:207-17

Ederhy, Stéphane ; Devos, Perrine ; Cohen, Ariel ; et al
From cardio-oncology to cardio-onco-pharmacology: Towards a multidisciplinary approach in the understanding and management of cardiotoxicity
Therapie 2022;77:197-206

dabrafenib-Tafinlar

Khouri, Charles ; Mahé, Julien ; Caquelin, Laura ; et al
Pharmacology and pharmacovigilance of protein kinase inhibitors
Therapie 2022;77:207-17

Figure 17 : La newsletter : L'Info Comme on l'M

1.3.2 PISTES DEJA EXPLOREES EN VUE D'UNE INDEXATION SEMI-AUTOMATIQUE

1.3.2.1.1 CUREBOT®[18]

Ma tutrice de stage, avait exploré des solutions pour optimiser le *workflow* de la documentation de l'ICM, elle avait repéré l'outil de veille Curebot®. Ayant une expérience de l'outil grâce à ma formation à l'INTD, où nous avons pu expérimenter l'outil plusieurs mois sur des projets de veille, je l'ai assistée lors de l'entretien que nous avons eu avec un de leur représentant commercial. Notre manager était également présent. Voici ci-dessous le compte-rendu de l'entretien qu'elle a gracieusement réalisé à ma demande.

Compte rendu

- **Objectif au départ :**

Recherche d'un outil permettant d'effectuer une détection automatique de plusieurs mots préalablement identifiés, dans une page web de sommaire électronique de revue (ou TOC « Table of Contents »). Ces mots correspondraient aux noms des molécules de médicaments commercialisés par Novartis, et dont nous devons surveiller toute nouvelle publication française dans nos abonnements souscrits *online*. Actuellement, la collecte de documents se fait titre par titre de revues, et article par article.

- **Prise de connaissance de Curebot® :**

Lors de ma présence au salon professionnel Documation du 24 mars2022, suite à mes visites sur les stands des exposants. Curebot® me paraissait répondre à notre besoin, car ils développent un outil automatique de collecte des ressources, et que peut-être ils auraient pu adapter sa technologie à nos sources spécifiques à surveiller.

- **Démonstration :**

Suite à la démo de l'outil, le 1er juillet 2022 par le dirigeant de Curebot®, et à nos discussions, il s'avère qu'ils sont spécialisés dans la mise à disposition de plateformes de veille collaborative, avec un catalogue de sources prédéfini avec partage et mutualisation des veilles de chaque collaborateur. Cela aboutit à un *dashboard* personnalisé avec les informations

[18] https://espritscollaboratifs.fr/curebot/ Consulté le 22/08/2022

collectées, mais aussi à une newsletter pouvant être partagée collectivement, sur un thème en particulier.

- **Freins et problématiques :**
    - La mutualisation des veilles est intéressante sur le principe, mais je pense que cela n'est pas applicable avec les sources que nous surveillons.
    - Gestion des accès aux pages web des éditeurs avec leur contenu html, PDF des articles. Quid de la gestion des mots de passe, et adresses IP ?
    - L'outil ne détectera pas les mots recherchés (noms de molécules) dans les tableaux, graphes, infographies... possiblement présents dans les publications.
    - Limite de propriété intellectuelle concernant l'utilisation de ces pages web éditeurs dans ce processus de lecture et de recherche de mots par une technologie d'IA.
    - Il faut aller demander les droits/autorisation à chaque éditeur pour utiliser ses pages web dans le *sourcing* et *crawling* d'un tel outil. Déclaration et régularisation contractuelle à faire obligatoirement pour chaque éditeur. Et inconnu sur l'attitude des éditeurs de presse par rapport à ça : Acceptation ? Coût supplémentaire ?

Je retiens de ce compte-rendu, qu'au-delà des limitation propre à l'outil lui-même, les freins juridiques, les limitations techniques et les restrictions des plateformes des éditeurs rendent impossible en l'état des choses le déploiement d'une solution entièrement automatisée qui s'étendrait de la veille à l'indexation dans IVAN. Or cette absence d'une solution de curation de la veille automatisée déçoit les espoirs de résorber le temps consacré à cette tâche de veille, qui comme nous l'avons montré en introduction est conséquente.

1.3.2.1.2 MENDELEY® [19]

[19] https://www.mendeley.com/?interaction_required=true Consulté le 22/08/2022

La collègue, en charge de la saisie des métadonnées d'indexation dans la base, utilise l'outil de références bibliographiques Mendeley® pour automatiser la saisie des champs de références bibliographiques. Elle charge les PDF des articles, que les différentes documentalistes lui soumettent, en *batch*[20] dans l'outil. Puis après correction, elle fait un import de ces métadonnées dans IVAN. Une fonction d'importation automatique de ces champs, à partir de l'outil Mendeley®, a été prévue, à cet effet, dans l'architecture de la base.

Cependant cela pose un bon nombre de problèmes :

- Le premier pour un néophyte est que la base a été conçue de telle manière qu'une variation (erreur humaine) dans les intitulés (différence de casse, formulation différente etc.) créée automatiquement une nouvelle valeur et indexe l'article sous celle-ci. Par exemple, bon nombre d'articles appartenant à la même revue sont indexés sous des références différentes. Et avec un import depuis Mendeley® sans ressaisie postérieure sous la bonne référence l'erreur est systématique. Ce qui rend l'outil inopérant.

- Le deuxième problème est que l'ICM utilise une norme qui lui est propre pour les références bibliographiques, ce qui rend une ressaisie, en ces termes, nécessaire.

- Le troisième problème est inhérent à l'outil lui-même : son algorithme n'est pas assez robuste et génère continuellement des erreurs de référencement (exemple : souvent le titre de l'article ou de la revue n'est pas le bon), et quand la pagination ou le statut de l'article est « exotique » (par exemple s'il est en prépublication (« *Advance Online »)* : là on est dans l'improvisation fantaisiste totale…

Le fait est que l'extraction automatisée des métadonnées est l'un de ces problèmes de l'IA qui semble très facile à résoudre mais qui est en fait assez difficile. Étant donné un article de recherche, qui a été bien formaté par un éditeur, il est normalement facile de repérer les métadonnées clés telles que son titre, les auteurs, où il a été publié et quand il a été publié. Le fait que les éditeurs utilisent une gamme variée de mises en page et de paramètres de type pour les articles n'est pas un problème pour les lecteurs humains. Or je les cite, (eux aussi sont capables d'ironie !) : « Si vous souhaitez qu'il [Mendeley®] extraie des métadonnées parfaites afin que vous puissiez les utiliser pour générer des citations à partir de Mendeley® Desktop sans avoir à effectuer de corrections manuelles, vous avez des exigences de qualité assez élevées. »[21]

[20] Un "batch" est en informatique le traitement de données par lots.

[21] https://krisjack.wordpress.com/2015/03/12/how-well-does-mendeleys-metadata-extraction-work/ Consulté le 22/08/2022

2 ELEMENTS DE CONTEXTE POUR MIEUX APPREHENDER CE QUE RECOUVRE UNE AUTOMATISATION DE L'INDEXATION.

2.1 CADRE REGLEMENTAIRE

2.1.1 L'ARTIFICIAL INTELLIGENCE ACT EST UNE PROPOSITION LEGISLATIVE PUBLIEE PAR LA COMMISSION EUROPEENNE LE 21 AVRIL 2021.[22]

« Le projet de règlement s'applique à un ensemble d'acteurs indépendamment de leur caractère public ou privé, ou de la nature de leur activité. Il vise donc autant les entreprises, les professions libérales, les associations, que les administrations ».[23]

Les acteurs sont caractérisés par leur rôle, leur fonction dans « la chaine d'approvisionnement du système d'IA » [24]:

1. Le « fournisseur » : « la personne qui le met sur le marché ou en service »

2. L'« importateur » c'est un « fournisseur » mais pour le compte d'une personne établie en dehors de l'UE.

3. Le « distributeur » qui n'appartient pas aux deux catégories précédentes qui met le système d'IA à « disposition sur le marché de l'Union sans altérer ses propriétés »

4. « L'utilisateur » : « la personne qui utilise le système d'IA sous son autorité dans le cadre d'une activité professionnelle. ».

Dans le cadre de Novartis France, l'entreprise serait considérée comme un « fournisseur » si elle mettait en service un système d'IA pour automatiser sa base de connaissance IVAN pour le compte de l'ICM à des fins commerciales et un « importateur » si elle décidait de l'exporter dans les autres filiales de Novartis au Global, toujours à des fins commerciales. « Un distributeur » si elle décidait de faire développer ce système par le Global puis qu'elle le diffusait aux autres filiales européennes de Novartis, à des fin commerciales. Cependant, si elle utilise le système à des fins de recherche ou de logistique, ce qui semble la finalité la plus évidente dans notre cas, elle sera considérée comme utilisateur ». Cet enchevêtrement peut

[22] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066 Consulté le 21/08/2022

[23] Petel, A. L'enjeu du champ d'application matériel : la définition de l'IA *in* Dossier : L'intelligence artificielle. I2Dn°1, juillet 2022.114p. (dir. Chartron G.). p. 24-25

[24] Ibid.

paraître un peu tirer par le cheveux, mais il laisse entrevoir que selon ce cadre réglementaire, une entité comme Novartis peut cumuler ces qualifications.

Une contrainte juridique proportionnée à quatre niveaux de risque[25].

1. Le premier niveau : L'atteinte aux règles et aux valeurs de l'Union Européenne. Soit une IA qui porterait un préjudice psychologique ou physique ou qui exploiterait la vulnérabilité d'une personne ou d'un groupe de personne. Par exemple, une IA permettant le crédit social est prohibée.

2. Le deuxième niveau de risque : « Les système d'IA à haut risque » entrainant un risque pour la santé, la sécurité ou les fondamentaux des individus.

Ces risques sont qualifiés par l'application des règles européennes sectorielle, listées dans l'annexe n°2 du projet de règlement, ou bien, à partir de la finalité d'utilisation du système d'IA (cf. annexe n°3).

2.1.2 COMMENT UTILISER DU CONTENU PROTEGE PAR DES DROITS D'AUTEUR POUR L'APPRENTISSAGE AUTOMATIQUE[26]

Malgré les inquiétudes concernant les utilisations de l'apprentissage automatique dans l'UE qui augmentent depuis un certain temps, ce n'est que récemment que les États membres ont commencé à adopter des exceptions similaires au droit d'auteur. Le Royaume-Uni a d'abord autorisé la reproduction non autorisée d'œuvres protégées par le droit d'auteur à des fins non commerciales de *Text and Data Mining* (TDM). La France, l'Allemagne et l'Estonie ont ensuite emboîté le pas. TDM est un terme général couvrant diverses méthodes d'analyse informatique de l'information qui incluent également l'apprentissage automatique et l'IA[27]. Voici sa définition par l'Union Européenne :

« toute technique d'analyse automatisée visant à analyser des textes et des données sous une forme numérique afin d'en dégager des informations, ce qui comprend, à titre non exhaustif, des constantes, des tendances et des corrélations»[28].

Lorsque les décideurs politiques européens ont commencé à réaliser l'importance de l'accès aux données pour le développement de l'IA dans l'UE, ils ont commencé à proposer des modifications des règles de l'UE en matière de droit d'auteur qui obligeraient chaque État

[25] Ibid.

[26] https://valohai.com/blog/copyright-laws-and-machine-learning/ Consulté le 21/08/2022

[27] https://valohai.com/blog/copyright-laws-and-machine-learning/ Consulté le 14/11/2022

[28] https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32019L0790&from=FR Consulté le 17/11/2022

membre à adopter les exceptions TDM correspondantes. Selon le dernier texte en date[29], l'exception permet à chacun de miner (pour *datamining* ou fouille de texte) du contenu auquel il a déjà accès. Cette directive a été transposé dans le droit français, par l'Ordonnance n° 2021-1518 du 24 novembre 2021[30]. Elle stipule que : « III. Sans préjudice des dispositions du II, des copies ou reproductions numériques d'œuvres auxquelles il a été accédé de manière licite peuvent être réalisées en vue de fouilles de textes et de données menées à bien par toute personne, quelle que soit la finalité de la fouille, sauf si l'auteur s'y est opposé de manière appropriée, notamment par des procédés lisibles par machine pour les contenus mis à la disposition du public en ligne. Les copies et reproductions sont stockées avec un niveau de sécurité approprié puis détruites à l'issue de la fouille de textes et de données. ». Au regard de ce qui nous intéresse c'est la phrase « quelle que soit la finalité de la fouille qui est déterminante.

Il est important de noter que les titulaires de droits qui avait jusqu'à très récemment généralement toujours le droit de restreindre l'utilisation de leurs œuvres à des fins de *datamining*, à l'exception de l'utilisation par des instituts de recherche à but non lucratif. En d'autres termes, seuls les instituts de recherche avaient le droit illimité d'exploiter le contenu protégé par le droit d'auteur, tandis que les autres acteurs devaient toujours respecter le choix de non-participation du titulaire des droits. Cette limitation visait à protéger les intérêts des éditeurs qui, tout en facturant aux abonnés un « accès en lecture », souhaitent néanmoins se réserver le droit de leur facturer séparément le droit de faire du *datamining*. Mais cette limitation n'était pas gravée dans le marbre. Et le Décret no°2022-928 du 23 juin 2022 portant modification du code de la propriété intellectuelle et complétant la transposition de la directive 2019/790 du Parlement européen et du Conseil du 17 avril 2019 sur le droit d'auteur et les droits voisins dans le marché unique numérique et modifiant les directives 96/9/CE et 2001/29/CE, prévoit comme contrainte seulement que : « Les personnes effectuant une fouille de textes et de données dans les conditions mentionnées au III de l'article L. 122-5-3 fournissent aux titulaires de droits d'auteur, à la demande de ceux-ci, tous documents et justificatifs permettant d'établir que les copies et reproductions numériques effectuées lors d'une fouille de textes et de données sont stockées avec un niveau de sécurité approprié et qu'elles ont été détruites à l'issue de la fouille de textes et de données ». En respect de cette condition, entraîner un modèle sur les articles dont dispose légalement l'ICM semble tout à fait possible !

2.2 CONTEXTE MESH

Les dernières années ont vu une augmentation rapide du nombre d'articles scientifiques dans le domaine biomédical. Ces documents sont pour la plupart disponibles et facilement accessibles sous forme électronique. La principale base de données bibliographiques mondiale est américaine. Elle regroupe la littérature relative aux sciences biologiques et biomédicales.

[29] https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=celex%3A32019L0790 Consulté le 14/11/2022

[30] https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000044362034 Consulté le 14/11/2022

Elle s'appelle MEDLINE (pour *Medical Literature Analysis and Retrieval System Online*) et sa version publique : PubMed, (pour PUBLIC MEDLINE). Elle est administrée et mise à jour par la bibliothèque étasunienne de médecine : la National Library of Medicine (NLM).

La base MEDLINE couvre 11 domaines :

Biochimie, biologie, médecine clinique, économie, éthique, odontologie, pharmacologie, psychiatrie, santé publique, toxicologie, médecine vétérinaire.

Un nombre croissant de références conduisent au texte intégral dans PubMed Central (PMC).[31]

Elle a plus de 28 millions de références bibliographiques provenant de 5 200 revues dans 40 langues.[32]

PubMed Central (PMC) contient les archives en texte intégral qui comprend des articles de journaux sélectionnés par la NLM, principalement de la littérature anglosaxonne. **Beaucoup d'articles provenant de revues francophones n'y sont pas indexés.[33]**

La classification de cette littérature biomédicale est réalisée grâce à un système de métadonnées médicales en langue anglaise fondé sur l'indexation d'articles en sciences de la vie : les termes MeSH (pour *Medical Subject Headings*).

Depuis le milieu des années 1960. Cette indexation se base, ainsi, sur ce vaste thésaurus : MeSH. C'est un vocabulaire contrôlé complet, qui a été développé et maintenu par la NLM. Par exemple, plus de 29 000 termes MeSH sont utilisés pour étiqueter plusieurs sujets de chaque résumé scientifique trouvé dans la base de données PubMed.

La force de PubMed réside dans l'utilisation du MeSH qui permet d'élaborer des requêtes précises (même si certains concepts très récents n'ayant pas encore d'entrée dans le MeSH ne sont pas requêtable à partir du MeSH).

Auparavant, ces métadonnées étaient attribuées manuellement, cette tâche est toujours en partie, actuellement, effectuée par des experts humains au NLM pour des milliers d'*abstracts* (résumés) par an. Chaque article biomédical publié est ensuite enregistré dans PubMed pour faciliter la récupération les informations pertinentes. L'indexation MeSH précise des documents est cruciale pour que les chercheurs biomédicaux puissent découvrir et diffuser de nouvelles

[31]André-Bourget, M.L. Intelligence artificielle et curation humaine : Medline 2022, projet d'indexation automatisée mené par la National Library of Medicine. *In* Dossier : L'intelligence artificielle, n°1 juillet 2022. 114p. (dir. Chartron G.) ; p.8

[32] Ibid.

[33] Ibid.

connaissances. Or, avec la croissance rapide de la base de données PubMed, l'indexation de documents biomédicaux à grande échelle est devenue de plus en plus importante.

La NLM a donc souhaité passer à l'indexation automatisée. Plusieurs travaux visant à automatiser le processus d'attribution des termes MeSH aux articles dans PubMed ont été expérimentés.

Ainsi Depuis 2002, la NLM utilise des algorithmes de traitement du langage naturel pour aider les indexeurs en fournissant des recommandations de termes MeSH.

Ces dernières années, la bibliothèque s'est de plus en plus appuyée sur l'assistance de l'IA et l'automatisation pour organiser plus efficacement la littérature biomédicale, et elle travaille continuellement à l'amélioration de la qualité de ces recommandations.

L'annotation manuelle par des curateurs humains est coûteuse en temps et en argent.

Un système de calcul qui peut aider à l'indexation est donc très précieux.

Depuis 2022, avec son outil MTI-a, La NLM propose une nouvelle approche de classement des textes et d'attributions des métadonnées, grâce au *Machine Learning*, qui permet une indexation, MeSH : « 100% automatisée ». En réalité Il s'agit d'un système hybride : une indexation entièrement automatisée, avec une curation humaine. Les indexeurs continuent d'apporter leur contribution, en perfectionnant les algorithmes d'indexation automatique et en exerçant un rôle de contrôle et d'assurance qualité. Depuis mi-2022, toutes les références sont censées être indexées via MTIA avec l'aide de cette curation humaine.[34]

La NLM fait en sorte qu'il y'ait une amélioration continue de son algorithme d'indexation automatique.

L'indexation MeSH est un défi, une tâche pour l'apprentissage automatique, car il doit attribuer plusieurs *labels* (étiquettes) pour chaque article à partir d'une liste de vocabulaire contrôlé, hiérarchisé, c'est-à-dire d'un thésaurus d'une taille extrême.

Si on généralise, l'indexation MeSH renvoie au problème d'attribution des *labels* les plus pertinents d'un document biomédical donné parmi un ensemble extrêmement large de termes MeSH.

Lors du développement de systèmes d'indexation MeSH supervisés, la disponibilité d'un corpus de textes annotés à grande échelle est souhaitable. Un vaste corpus accessible au public qui permet une évaluation et une comparaison robustes de divers systèmes est important pour la communauté de la recherche.

[34] Medline 2022. Initiative: transition to automated indexing. NLM Tech Bull, 2021, (443):e5. https://www.nlm.nih.gov/pubs/techbull/nd21/nd21_medline_2022.html Consulté le 21/082022

L'algorithme de MTIA n'est à l'heure actuelle pas rendu public. Mais plusieurs corpus d'indexation MeSH annotés à grande échelle ont été rendus disponibles. Comme MeSHup qui contient 1 342 667 articles en texte intégral en anglais, ainsi que ses *labels* et métadonnées MeSH associées. Ce corpus peut nous servir de première *baseline*, dans une première approche.

A noter qu'il existe un MeSH bilingue anglais-français, traduit par l'INSERM (il est en cours de refonte, la dernière version mise à jour date donc de 2019[35]

2.3 TAL[36]

L'intelligence artificielle et l'apprentissage machine (*Machine Learning*, ML) se nourrissent de données (*Data*) plus ou moins massives, afin de dériver des modèles statistiques capables de reproduire une fonction réalisée habituellement par un humain. Afin d'effectuer des analyses prédictives, prospectives ou prescriptives.

Le Traitement du langage naturel (TAL) est un ensemble de technologies issues historiquement de la sémantique et de la linguistique, plus récemment des technologies d'apprentissage (*Machine Learning*) et d'apprentissage profond (*Deep Learning*). Suivant les approches et contextes, les intervenants peuvent utiliser l'une ou l'autre de ces technologies ou une combinaison des deux (approche hybride).

Concernant la branche de l'IA, du traitement automatique du langage naturel (TAL ou TALN ou encore NLP en anglais pour *Natural Langage Processing*) : deux mondes très différents se partagent le domaine.

L'IA historique c'est-à-dire l'IA symbolique avec ses modèles à règles et l'IA connexionniste basée sur l'apprentissage profond. Même si ces deux catégories peuvent être combinées, nous proposerons des solutions basées seulement sur cette deuxième catégorie pour

[35] https://mesh.inserm.fr/FrenchMesh/. Consulté le 21/08/2022

[36] NB : Ceci est une synthèse adaptée du numéro d'I2D (L'intelligence artificielle, I2D n°1 juillet 2022,114p. (dir. Chartron G.) qui me sert de référence. Pour éviter une paraphrase inutile et malhonnête, pour la clarté et la finesse de leurs formulations la quasi-totalité de ces dernières ont été restituées telles quelles, quelques informations complémentaires et raccords ont été ajoutés pour plus de précisions. Ce numéro a réuni plusieurs experts, leurs propos sont ici mélangés, je ne les ai pas cités nommément pour chaque phrase emprunté (en note de bas de page) pour ne pas alourdir le document, mais j'invite le lecteur curieux à lire ce numéro dans son intégralité car il est d'une grande qualité.

des questions de performances et parce que la première nécessite un haut niveau d'expertise, et qu'en toute sincérité est largement dépassée aujourd'hui.

À titre d'exemple, on se souvient du saut qualitatif de Google Translate le jour où, en 2016, la plateforme a basculé son service sur des méthodes d'apprentissage profond.

IA connexionniste est fondée sur l'apprentissage d'une tâche à partir d'exemple d'entrée/sortie de cette tâche.

Depuis son apparition dans les années 50, l'intelligence artificielle a énormément progressé. La capacité accrue du calcul machine, l'apprentissage automatique, perfectionné avec des algorithmes puissants, et optimisé par des approches combinatoires tirant profit des technologies de la langue, l'ont portée à un stade d'industrialisation avancé, si bien qu'elle est en passe d'opérer une véritable synergie entre l'intelligence créative de l'être humain et les capacités de traitement de l'information ou de la communication multimédia.

Pour des tâches qui étaient réservées à la seule approche humaine elle est désormais présente à toutes les étapes du cycle de l'information (classification, génération, prédiction...). L'IA révolutionne les pratiques et les métiers documentaires : elle ouvre de nouvelles possibilités pour gérer et valoriser les contenus, qui sont autant de défis.

Un saut technologique nous a fait basculer dans une nouvelle ère : celle des techniques statistiques d'apprentissage machine exploitant des données massives, qu'elles concernent des documents textuels, multimédia ou des données factuelles.

En pratique, ce sont ces données massives qui caractérisent les usages de l'IA. Ce saut quantitatif dans le volume de données constitue à la fois la condition et la portée de ces algorithmes. Les réseaux de neurones ont en effet besoin de s'être (ou d'avoir été) entraînés sur de grands jeux de données. Plus les données sont massives, plus les prédictions seront justes.

L'information et les données peuvent y être considérées comme le « carburant » des algorithmes de l'IA, dans la diversité de leur complexité.

Dans le secteur de l'info-doc (gestion des contenus, des données et des documents), l'IA est aujourd'hui perçue comme une opportunité pour optimiser les processus et les flux de tâches existants liés à la production, à la gestion, à la valorisation et à l'usage de la donnée (collecte, agrégation, enrichissement, analyse, livraison, etc.). Pour des traitements de masse à grande échelle qui ne pourraient être réalisés par la seule approche humaine et engendre des bouleversements sur la chaîne de valeur.

Il devient alors essentiel de prendre la mesure des impacts de l'intelligence artificielle sur le cycle de l'information que ce soit au niveau de la collecte, du traitement, de l'analyse et de la diffusion des contenus textuels et multimédia.

Le principe fondamental de ces algorithmes : ces derniers n'opèrent pas sur des objets préalablement modélisés informatiquement. Au contraire, ils façonnent, à partir d'un grand jeu de données, leurs propres modèles. Plus exactement, ces algorithmes repèrent dans les données des motifs (*patterns*) sur lesquels ils peuvent opérer des classifications. Si l'on veut préciser encore une tâche fondamentale, on dira que ces algorithmes, après avoir entraîné de manière autonome ou supervisée un réseau de neurones à identifier des motifs au sein des données, vont être en mesure de donner une prédiction de classement sur une nouvelle donnée. On peut visualiser ces motifs comme des empreintes, des chemins décisionnels au sein des réseaux de neurones. Ces derniers sont alors capables d'opérer des tâches dites « de bas niveau ». Mais aussi vers des tâches de plus haut niveau initialement réservées à l'humain : analyse terminologique, syntaxique ou sémantique, identification d'entités nommées, calcul de distance entre différentes unités documentaires (concepts, paragraphes, articles, corpus), modélisation d'un texte en un graphe de connaissances, rédaction de résumés, ou pourquoi pas d'articles entiers, etc. Ces applications laissent entrevoir de nouvelles possibilités de requête au sein de corpus textuels.

L'accès et l'analyse d'information technique est également optimisé qu'il s'agisse de littérature scientifique ou de documentations techniques. Les moteurs de recherches peuvent être spécialisés et contextualisés par l'ajout de base de connaissance dont la constitution est assistée par l'IA.

Plus spécifiquement, l'apprentissage profond appliqué au TAL va permettre la génération de métadonnées, d'extraction d'information et de relations. Mais aussi de multiplier les dimensions de classement de l'information, de recherche et d'accès guidé à l'information, d''analyse de grandes masses de textes ou documents et d'exploration de corpus.

Plus techniquement le TAL permet d'identifier :

- Les entités nommées : noms de personnes, sociétés, organisations, lieux, produits...
- Les concepts ou thématiques (i.e. *topics*) qui définissent les mots ou groupes de mots ayant une signification par rapport aux sujets couverts au sein du document ;
- Les relations, c'est-à-dire les liens entre deux entités ou entre une entité et un concept (par exemple une société est partenaire d'une société, ou telle société lance un nouveau produit dans tel domaine...).
- Permet de gérer et d'analyser de grandes quantités de documents, de mettre à jour des vocabulaires métiers et des bases de connaissance, et ainsi d'économiser un temps considérable. Les vocabulaires structurés (lexiques, dictionnaires, thésaurus, taxonomies...)

Un bon exemple est le thésaurus médical MeSH

Ces vocabulaires doivent être constamment mis à jour car les domaines qu'ils couvrent évoluent. Cette mise à jour peut s'avérer une tâche lourde et fastidieuse surtout lorsqu'il y a des dizaines de milliers de concepts à gérer ou plus encore !

Pour un professionnel de l'information, il est primordial de se tenir au fait des informations pertinentes dans les domaines biomédicaux. Une évolution constante des concepts ou notions biomédicaux, les concepts émergents, les nouveaux acronymes, nouveaux termes utilisés, le droit de la propriété intellectuelle poussent à une veille permanente. Suivre ces questions avec des approches manuelles devient rapidement impossible surtout si les destinataires de l'*Info comme on l'M* s'intéressent à d'autres sujets tout aussi complexes. Une assistance devient indispensable, utile, à l'activité de l'ICM.

3 QUALIFICATION DE LA PROBLEMATIQUE TECHNIQUE

3.1 FONDEMENTS TECHNOLOGIQUES

Beaucoup d'informations, dans le monde, sont non structurées, le Traitement Automatique de la Langue ou TAL[37] est le champs de l'intelligence artificielle qui permet de comprendre les textes non structurés et d'en extraire des données. Plus largement le TAL, et le domaine qui permet aux ordinateurs de lire, comprendre et déduire du sens à partir des langages humains.

Historiquement le TAL est issu de la sémantique et de la linguistique et a donnée l'IA symbolique ou les modèles à règles pour reproduire une fonction réalisée habituellement par un humain. Ces systèmes experts très complexes, calqués sur le raisonnement humain, difficilement généralisables et peu opérants, ont été largement remplacés par le *Machine Learning* (apprentissage machine) capable, à partir de données massives, de dériver des modèles statistiques pour effectuer des analyses prédictives, prospectives ou prescriptives. Plus récemment, un technologie particulière de *Machine Learning*, le *Deep Learning* ou apprentissage profond, grâce aux progrès majeurs des capacités de calcul, plus précisément des cartes graphiques, : les GPU qui permettent de réaliser un bon nombre de calculs parallèles, a fini par dominer de manière quasi-hégémonique le champs de l'intelligence artificielle et il est à l'origine des percées exceptionnelles dans tous les domaines de l'IA depuis les années 2010.

Toutes ces techniques d'intelligence artificielle, au plus haut niveau d'abstraction comme elles sont conceptualisées dans le diagramme ci-dessous, ne différent guère d'un

[37] Dérivé de l'anglais NLP pour : *Natural Language Processing*

programme informatique traditionnel. Elles se nourrissent de données en amont, les digèrent dans un algorithme et produisent un résultat attendu. Voir le schéma ci-dessous[38] :

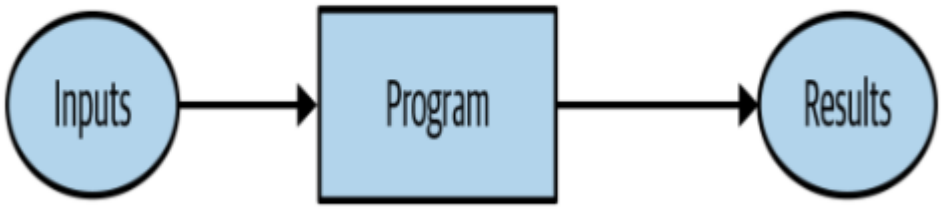


Figure 1-4. A traditional program

Figure 18 : Abstraction de ce qu'est un programme au plus haut niveau

Là où l'apprentissage profond diffère : c'est qu'il a pour architecture un réseau de neurone qui est une imitation directe, une transposition d'un réseau de neurones humains. On peut voir dans l'illustration suivante[39] le schéma d'un neurone humain et son abstraction informatique.

[38] *In* Howard, J., Gugger, S. Deep Learning for Coders with Fastai and PyTorch: AI Applications Without a PhD. O'Reilly Media, 2020

[39] Ibid.

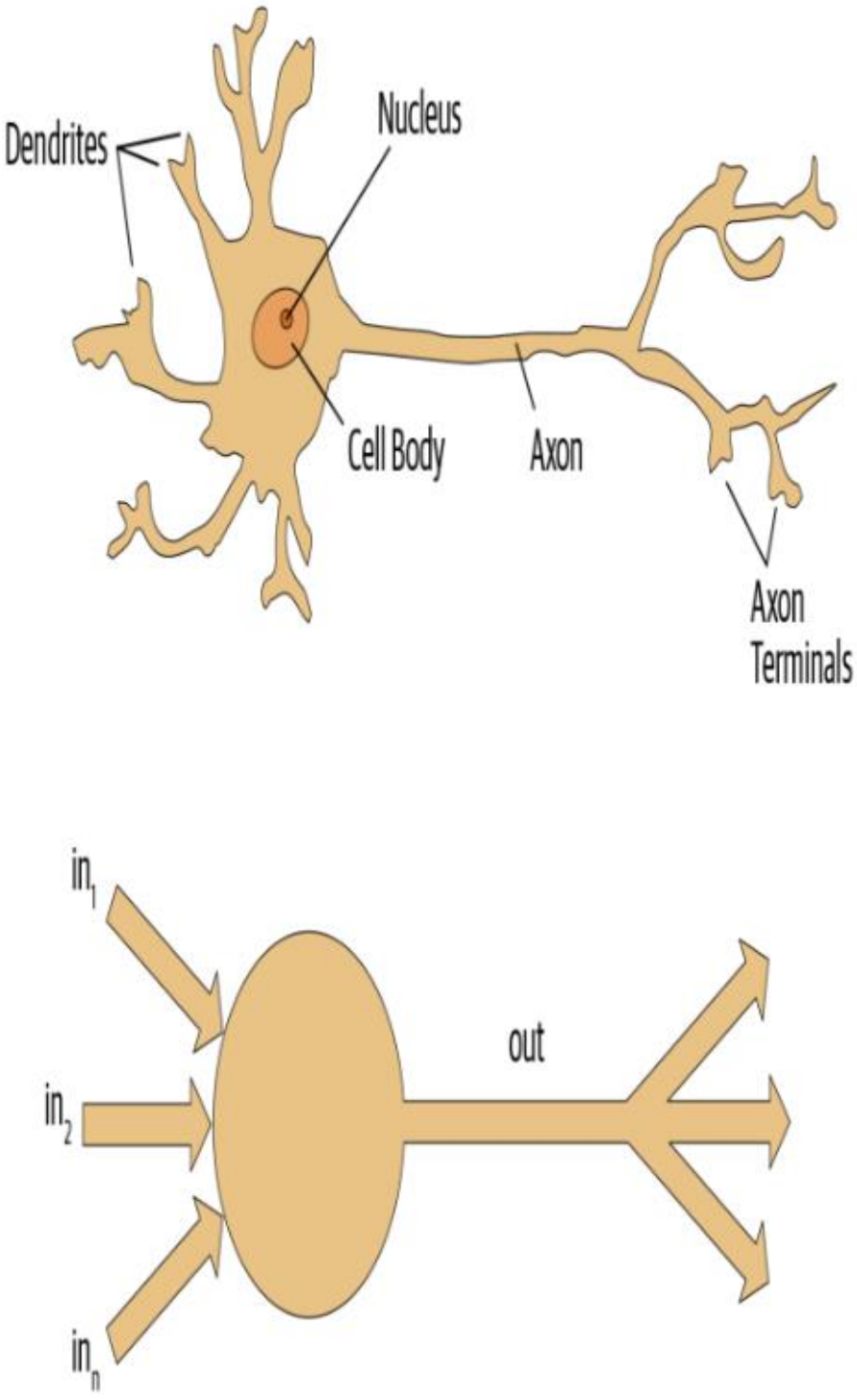


Figure 19 : Schéma d'un neurone biologique et d'un neurone artificielle

Un réseau de neurone est l'interconnexion en couches successives (*layers*) de neurones (ci-dessous[40] symbolisés par les cercles de couleurs). On parle d'apprentissage profond, puisque l'essentiel de la magie s'opère dans les couches profondes (« cachées ») du réseaux (les *hidden layers*). Plus le réseau est « *deep* », plus il a de couches cachées, plus il est en mesure de saisir la complexité des données en *inputs*.

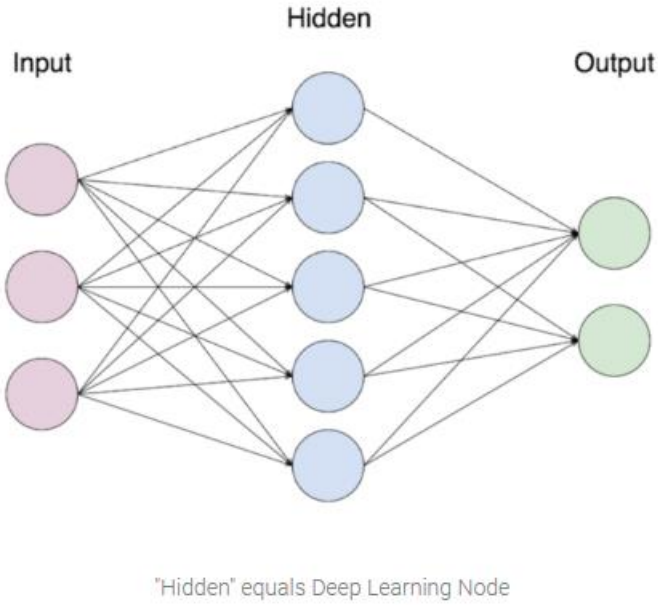


Figure 20 : Vue schématique d'un réseau de neurones

Alors qu'un programme ou un modèle à règle est défini à la main, un réseau de neurones est censé générer des représentations, des *patterns*, des motifs, en d'autres termes : de généraliser pour aboutir à une solution. C'est en ce sens qu'on peut parler d'apprentissage (*learning*).

Dans le diagramme ci-dessous[41] on a une comparaison du mode de fonctionnement des modèles utilisés en intelligence artificielle et leurs différences.

[40] In https://towardsdatascience.com/step-by-step-guide-to-building-your-own-neural-network-from-scratch-df64b1c5ab6e Consulté le 17/11/2022

[41] Goodfellow, I., Bengio, Y., Courville, A. Deep Learning. MIT Press, 2016

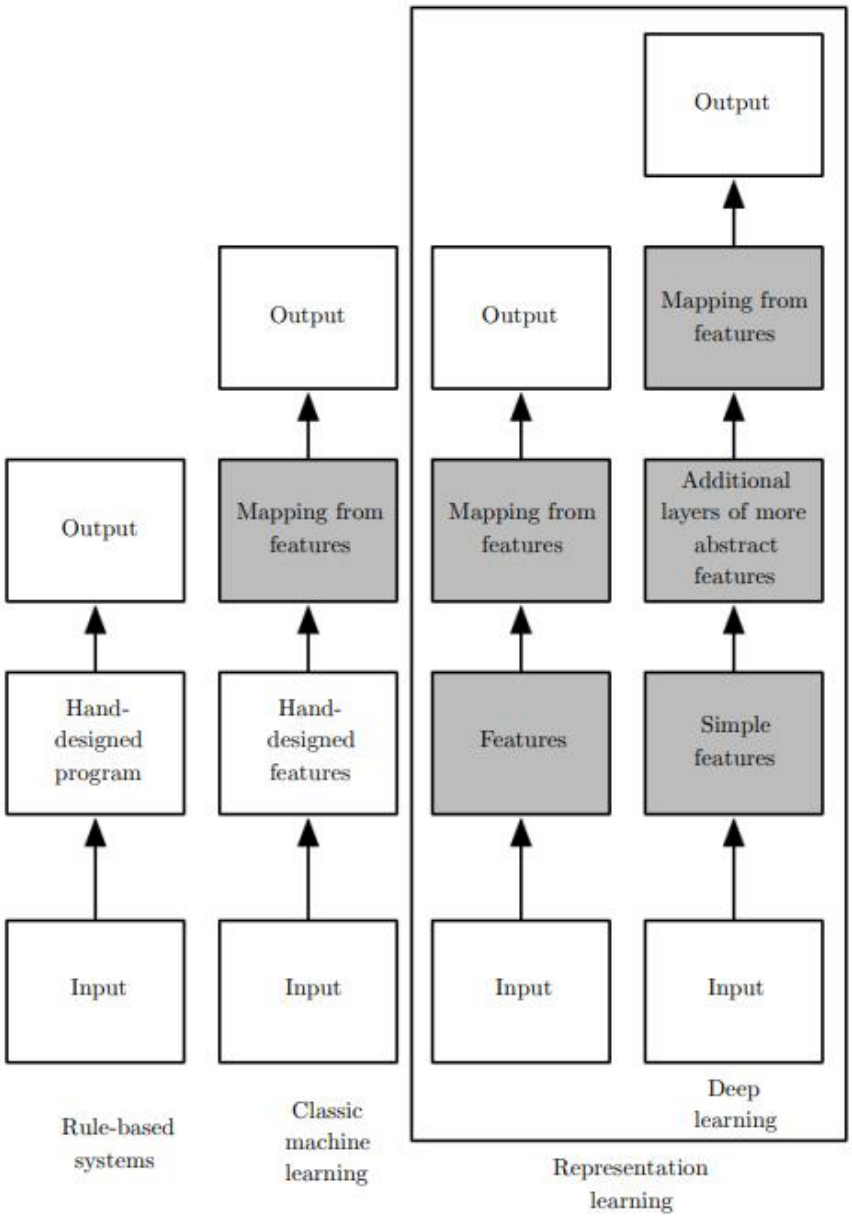


Figure 21 : Diagramme de différents types de modèles d'intelligence artificielle

Ici « *features* » peut être traduit par caractéristiques. Ainsi dans le champs qui nous intéresse, le TAL, un modèle qui est piloté par un réseau de neurone profond, va être en mesure (même si, en lui seul, il n'est pas suffisant, comme nous allons le voir plus loin) grâce à ses couches cachées, de déceler des caractéristiques

lexicales, grammaticales et sémantiques, de plus en plus abstraites, plus on avance en profondeur dans ses *layers*.[42]

Ci-dessous[43], la version mathématique d'un neurone artificielle.

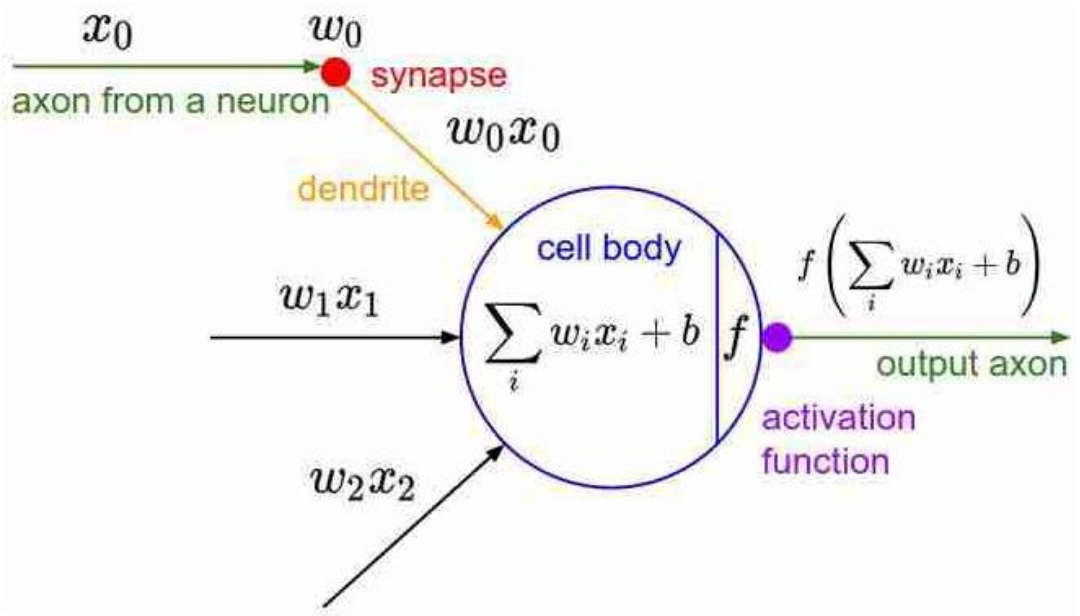


Figure 22 : Version mathématique d'un neurone artificielle

On a ici la structure fondamentale d'un réseau de neurones et ses différent paramètres.

Les paramètres ne sont que des chiffres. Ici X_n correspond aux données, aux *inputs* qui nourrissent le modèle en entrée. Ça peut être des chiffres représentant les valeurs des pixels d'une image dans le cadre de la reconnaissance d'image. Par exemple, des valeurs comprises entre 0 et 255 pour chaque pixel, ou son degré de luminosité : une nuance de gris allant de 0 pour le noir jusqu'à 255 pour le blanc. Mais ça peut tout aussi bien être des valeurs représentant les unités distinctives d'un spectre sonore pour la reconnaissance de la parole, donc des

[42] Il faut noter toutefois, qu'un réseau de neurone est particulièrement opaque dans son fonctionnement. C'est une véritable boite noire. Il est rare qu'un réseau de neurone ait une idée précise d'un concept (lexical, grammatical, ou sémantique) et donc on ne peut qu'avoir que des intuitions quant aux caractéristiques qu'il extrait des données. Elles sont rarement pures, c'est-à-dire qu'un réseau de neurones a rarement « les idées claires » comme l'entendait Descartes pour qualifier l'entendement humain.

[43] https://www.cs.ox.ac.uk/files/11720/Lecture11NLP3.pdf Consulté le 17/11/2022

phonèmes. Plus proche de notre problématique, ces valeurs peuvent être celles d'un *token* ou l'unité fondamentale d'une phrase : un mot ou une syllabe encodé sous forme d'un vecteur qui dans le cadre du *Word embbeding* encapsule sa position dans la phrase ainsi que sa sémantique, ses relations de proxémie (de quels termes, il est proche ou au contraire différent) avec les autres *tokens* et dans un vocabulaire donné, comme nous le détaillerons plus loin.

Ces *inputs* sont pondérés par d'autre paramètres : les poids, ici symbolisés par W_n. Ces poids sont associés à chaque valeur initiale et amplifient ces valeurs ou au contraire les minimisent. Chaque valeur associé à son poids est additionné dans un neurone qui est la somme pondérée de toutes ses valeurs, transformées et synthétisées par une fonction d'activation qui produit une nouvelle valeur comprise entre 0 et 1 et qui traduit l'état du neurone : plus la valeur est proche de 1 et plus le neurone est « actif ». C'est-à-dire qu'il a détecté positivement ou négativement la présence d'un *pattern* ou motif dans les *inputs*. Par exemple, dans le cadre de la reconnaissance d'image, si l'on considère la première couche comme étant les valeurs bruts des pixels d'une image, la seconde couches est constituée, exclusivement, de plusieurs neurones qui, respectivement, s'activent quand la somme pondérée des *inputs* de la premières couche correspondent à une forme élémentaire comme un *edge*, un bord spécifique, et suffisant pour être caractérisé. L'information va se propager vers la couche de neurones suivante et ainsi monter en abstraction.

Chaque couche de neurone va combiner *les patterns* des couches précédents en des motifs de plus en plus complexe, comme on peut le voir dans l'illustration ci-dessous[44], pour aboutir à la classification souhaitée (ici « *object identity* »).

[44] Goodfellow I., Bengio Y., Courville, A. Deep Learning. MIT Press, 2016

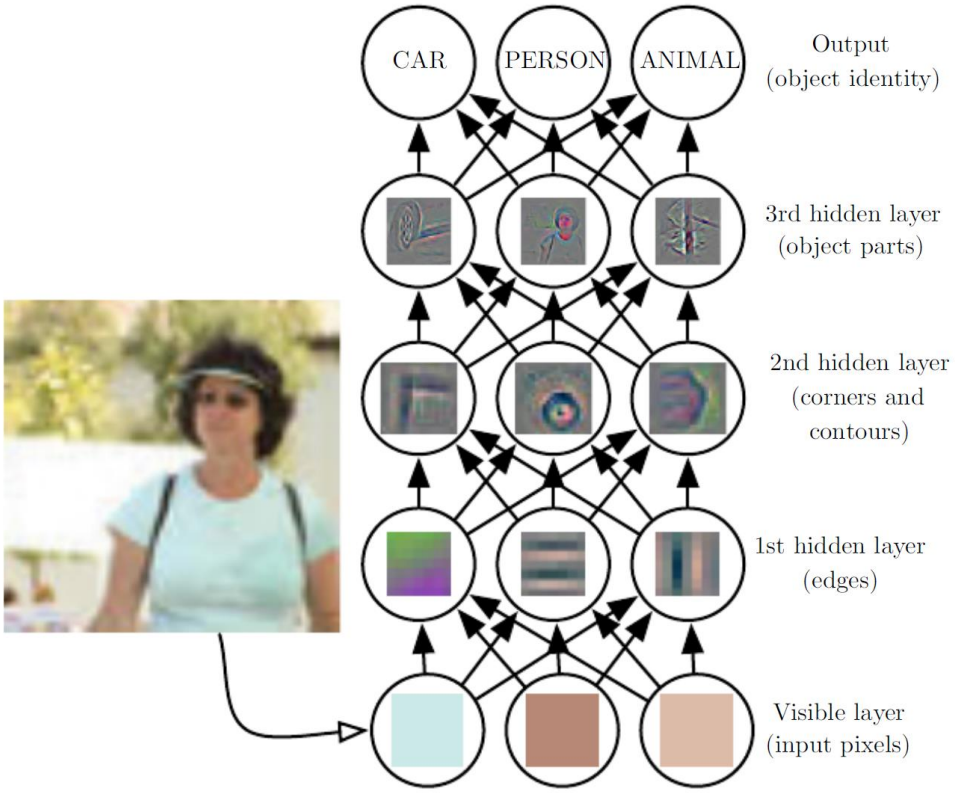


Figure 23 : Classification d'une image par réseau de neurones

Un modèle bien entraîné, ici, aboutira à la classification de cette image comme relevant d'une personne et non d'une voiture ou d'un animal.

On entend donc, ici, l'apprentissage profond comme l'ajustement des paramètres (les poids et les biais[45]) de chaque couche permettant l'aboutissement à cette classification.

Cet apprentissage profond est dans le cadre du *deep learning* : l'optimisation d'une fonction de coût par la technique de descente du gradient.

Une fonction de coût est une fonction composée qui à partir du calcul de la différence entre l'*input* et l'*output*, à la sortie du classifieur détermine l'amplitude de l'erreur. Ici dans notre exemple, si par exemple le réseau de neurone a classé l'image comme « voiture » et non

[45] Le biais est un nombre conséquent qui joue le rôle de seuil (en l'additionnant à la somme pondérée) au-dessus duquel le neurone va s'activer ou au contraire rester dormant. Cette technique n'est pas présente dans tous les modèles (Notamment dans les modèles récents que nous allons présenter plus avant dans le mémoire).

comme « personne », le réseau de neurone va mettre à jour ses paramètres (les poids et les biais) en les corrigeant pour qu'à la prochaine itération chaque image soit mieux classée. Il fait ça étape par étape en calculant le gradient de la fonction de coût. C'est-à-dire, les dérivées partiels, un vecteur compilant l'amplitude de modification nécessaire pour chaque paramètre pour maximiser globalement la fonction de coût. C'est le rôle du gradient. Pour l'apprentissage profond il faut en prendre la direction inverse, puisqu'on cherche, nous, à la minimiser, jusqu'à trouver un minimum local, le creux d'une vallée (cf. la modélisation ci-dessous[46]), l'optimum d'apprentissage de notre réseau de neurones (cf. la deuxième modélisation[47]).

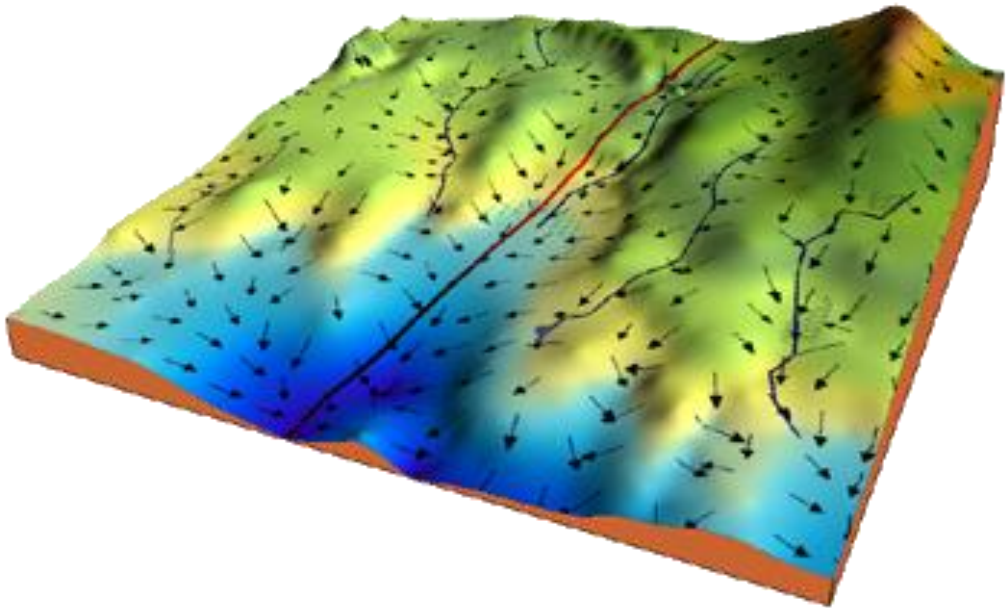


Figure 24: Modélisation imagée en vallée d'une descente de gradient

[46] https://ml-cheatsheet.readthedocs.io/en/latest/gradient_descent.html Consulté le 17/11/2022

[47] https://easyai.tech/en/ai-definition/gradient-descent/ Consulté le 17/11/2022

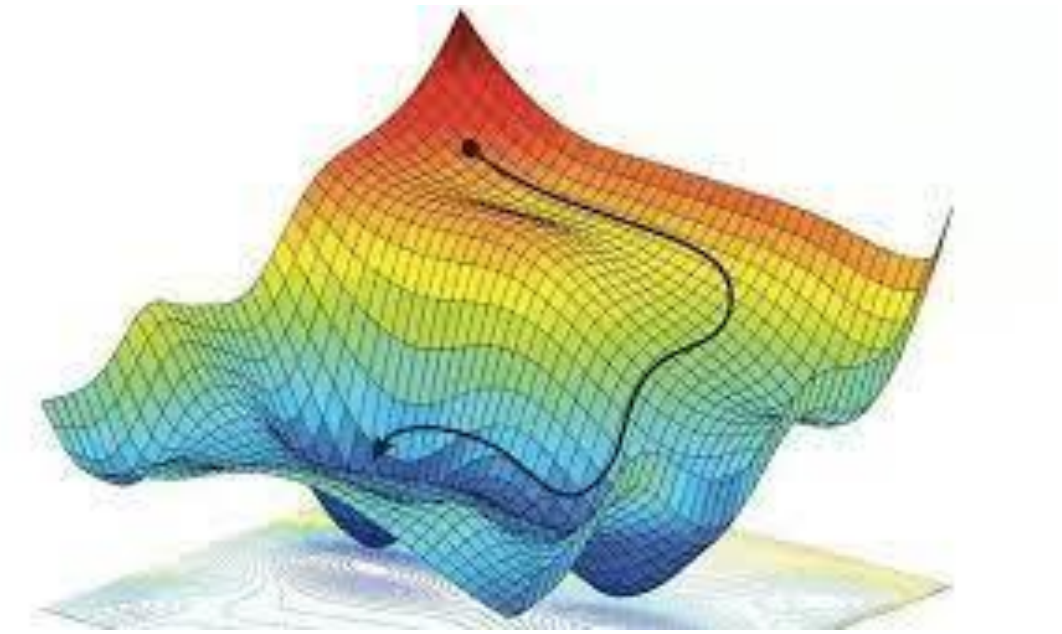


Figure 25 : Modélisation, mathématique, d'une descente de gradient

En théorie, on calcule le gradient moyen des fonctions de coût de toutes les itérations sur notre ensemble de données d'entraînement. Ainsi, idéalement la descente de gradient s'opère après avoir calculé le gradient moyen de toutes les fonctions de coût de la totalité des données d'entraînement, mis à jour après avoir passé en revue toutes les données dans une première itération. Puis le procédé est répété jusqu'à atteindre un minimum local. Mais en pratique cela nécessiterait trop de ressources de calcul et de mémoire étant donnée les larges quantités de données nécessaire pour un apprentissage suffisant. En effet, les modèles actuels les plus performants sont entraîné sur des milliards, voire des centaines de milliards de données. Donc, en réalité, on calcule le gradient sur un échantillon pour approximer le gradient global de manière stochastique, et on met à jour les paramètres régulièrement sans attendre d'avoir traversé l'ensemble des données d'entraînement.

Les algorithmes d'optimisations, sont des variations sur ce thème plus ou moins sophistiquées, il en existe un grand nombre, et des innovations sont fréquent, mais ils respectent tous cette idée de départ. Ce principe fondateur du *Deep Learning*, de différenciation d'une fonction de coût par la descente de gradient, s'appelle dans le jargon la rétropropagation, et c'est le mécanisme qui est au cœur de l'apprentissage profond et qui en fait son succès.

Si les algorithmes et les modèles évoluent à grand pas et remettent en question les acquis précédents, ce principe fondamental n'est jamais remis en cause.

La magie de l'apprentissage d'une machine dans le cadre du *Deep Learning* s'opère prosaïquement par un algorithme de rétropropagation qui mathématiquement est l'application de la « *chain rule* », théorème de l'Analyse qui permet de calculer la sensibilité d'un paramètre : quel ajustement est nécessaire, pour optimiser la fonction de coût à cet endroit, en relation avec tous les autres paramètres : dans quel direction il faut aller pour la minimiser.

Pour conclure, on peut dire qu'un modèle de réseau de neurone, n'est rien d'autre qu'une « grosse » fonction mathématique avec un très, très, grand nombres de paramètres. Elle nécessite, donc, beaucoup de capacités de calcul, mais peu de « jus de cerveau » pour sa mise en œuvre. Au contraire, d'un modèle à règle ou système expert qui est son strict opposé : beaucoup de « jus de cerveau » mais peu de capacités calcul.

Les *Word embeddings* ou plongement lexical, sont les représentations numériques, sous forme de vecteurs, des unités fondamentales d'un texte, soit les mots ou des « sous-mots » (les *tokens* de BERT, par exemple, comme nous le verrons plus tard). Plus exactement, chaque mot, d'une vocabulaire fixé préalablement, est relié à un vecteur dans un espace à N dimension. C'est-à-dire que ces vecteurs sont corrélés, contextualisés et donc porteur de sens.

En effets Les *Word embeddings* tentent de capturer la signification sémantique, contextuelle et syntaxique de chaque mot, dans le vocabulaire du corpus, en fonction de l'utilisation de ces mots dans les phrases. Les mots qui ont une signification sémantique et contextuelle similaire ont également des représentations vectorielles similaires, tandis qu'en même temps, chaque mot du vocabulaire aura un ensemble unique de représentations vectorielles.

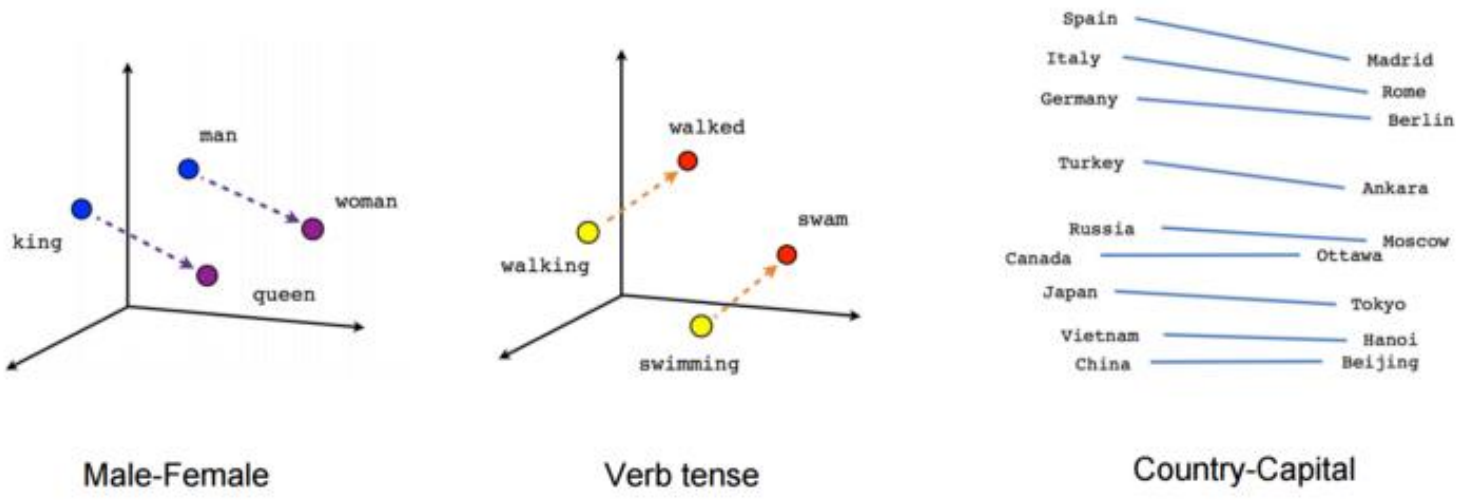


Figure 26 : Représentations graphiques de la similarité entre différents Word embeddings

L'image ci-dessus[48] montre des exemples de mots dans le vocabulaire ayant une signification contextuelle, sémantique et syntaxique similaire, représentés dans un espace vectoriel tridimensionnel. Dans le premier graphique, qui représente la relation de genre : les différences vectorielles entre les pairs de mots « roi » et reine » et « homme » et « femme » sont à peu près égales : les deux vecteurs sont parallèles. Ils sont porteurs d'une signification contextuelle similaire.

Globalement, un *Word embedding* arrive à découvrir des spécificités inhérentes à la langue comme la synonymie, l'analogie, les relations de généralisation/spécialisation (Les pays et leur capitale dans le troisième graphique) ou encore la morphologie (dans le deuxième graphique).

Si on reprend l'exemple du premier graphique, une telle transformation en vecteur permet de transposer une question comme « un homme est à une femme ce qu'un roi est à … ? » en une simple addition vectorielle : V(Roi) -V(Homme) + V(Femme) = V(Reine).

En effet, le calcul d'un plongement lexical s'appuie sur l'hypothèse distributionnelle de Harris : « un mot est défini par les mots qui l'entourent ».

Depuis 2018, un nouveau modèle de la langue a bouleversé le champs du TAL. C'est le modèle BERT qui utilise une nouvelle technique appelé Masked LM (MLM) : il masque au hasard des mots dans la phrase, puis il essaie de les prédire. Le masquage signifie que le modèle regarde dans les deux sens, de gauche à droite et de droite à gauche, il utilise le contexte complet de la phrase afin de prédire le mot masqué.

Le sens d'un mot est donc évalué en fonction de ses contextes dans le corpus d'apprentissage et non seulement de comment il s'écrit.

Avant BERT, pour les modèles contextuels unidirectionnels le mot « *bank* » en anglais aurait la même représentation, hors contexte, dans « *bank account* » (« compte bancaire ») et « *bank of river* » (« la berge de la rivière »). Parce que dans la phrase « *I accessed the bank account* », ces modèles auraient représenté « *bank* » en se basant seulement sur « *accessed* » mais pas sur « *account* ». Alors que BERT représente « *bank* » en utilisant à la fois son contexte précédant et suivant : « *I access the…account* ».

De plus, BERT adopte l'architecture, des révolutionnaires *transformer*s (cette architecture est à l'origine des avancées majeures du TAL ces dernières années). Un *transformer* fonctionne en effectuant un petit nombre constant d'étapes. À chaque étape, il applique un mécanisme d'attention pour comprendre les relations entre tous les mots d'une phrase, quelle que soit leur position respective. Par exemple, étant donné la phrase « *I arrived*

[48] https://www.tensorflow.org/tutorials/representation/word2vec Consulté le 17/11/2022

at the bank after crossing the river »[49] , pour déterminer que le mot « *bank* » fait référence à la rive d'une rivière et non à une institution financière, le *transformer* peut apprendre à faire immédiatement attention au mot « *river* » et prendre cette décision (de prendre en compte cette partie de la phrase pour déterminer ce sens précis de « *bank* ») en une seule étape.

BERT est ce qu'on appelle : *a pretrained model*, c'est-à-dire qu'on peut « transférer » ses *Word embedding* et les utiliser pour la tâche et le corpus qui nous intéresse (*transfer learning* en anglais) par le procédé du *fine tuning* ou réglage fin en français, un ajustement préalable. Ou dit autrement, c'est un processus qui prend un modèle qui a déjà été entraîné pour une tâche donnée, puis ajuste les paramètres ou ajuste le modèle pour lui faire effectuer une deuxième tâche proche.

Dans le TAL cela a été un véritable changement de paradigme. Car après avoir été entraîné sur une quantité astronomique de données textuelles, un modèle comme BERT peut être réutilisé et ajusté rapidement sur une tâche particulière qui ne nécessite qu'un petits nombre de données labélisées. En effet, il a déjà appris les *patterns* linguistiques génériques lors de sa phase de pré-entraînement.

La principale limite est que BERT a généré des *Words embeddings* pour un vocabulaire fixe de 30 000 *tokens* qui comprend les principaux mots de la langue anglaise, et les préfixes et les suffixes les plus utilisés à partir desquels il recompose les mots qui lui sont inconnus. Pour un domaine spécialisé comme le biomédical, il aura donc du mal à dégager des représentations et des relations pour des termes qui sortent du langage commun.

Il est alors nécessaire de pré-entraîner BERT, en partant de zéro, sur un corpus spécialisé (dans notre cas biomédical, ou dans la langue qui nous intéresse avant de le *finetuner* sur la tâche qui nous intéresse. Ici, la classification *multi-label*, d'articles scientifiques biomédicaux.

La classification *multi-label* est issue de l'étude du problème de catégorisation de texte, où chaque document peut appartenir simultanément à plusieurs sujets prédéfinis.

La classification *multi-label* des données textuelles est un problème important. Couramment, les exemples vont des articles de presse aux courriels. Par exemple, cela peut être utilisé pour trouver les genres auxquels appartient un film, sur la base du résumé de son intrigue.

Dans la classification *multi-label*, l'ensemble d'apprentissage est composé de textes, chacun associé à un ensemble de *labels*, et la tâche consiste à prédire les ensembles de *labels* de textes inconnus en analysant les données d'apprentissage avec des ensembles de *labels* connus.

La différence entre la classification multi-classe et la classification *multi-label* est que dans les problèmes multi-classe, les classes sont mutuellement exclusives, alors que pour les

[49] « Je suis arrivé à la berge après avoir traversé la rivière »

problèmes *multi-label*, chaque *label* représente une tâche de classification différente, mais les tâches sont en quelque sorte liées.

Par exemple, la classification multi-classes fait l'hypothèse que chaque échantillon est affecté à un et un seul *label* : un fruit peut être soit une pomme soit une poire mais pas les deux à la fois. Alors qu'un exemple de classification *multi-label* peut être qu'un texte peut porter sur n'importe lequel de ces sujets : religion, politique, finance ou éducation en même temps ou aucun de ceux-ci.

La plupart des méthodes de classification *multi-label* sont basées sur une combinaison de classifieurs binaires, qui sont entraînés séparément pour chaque *label*.

Toutefois, jusqu'à maintenant, beaucoup souffraient d'une précision et d'une efficacité modestes, avec seulement un succès limité dans l'utilisation pratique.

Or, récemment, il y'a eu des progrès dans la classification *multi-label* basée sur l'apprentissage automatique (*machine learning*) et plus particulièrement sur l'apprentissage profond. Des techniques de classification de textes médicaux ont été utilisées pour aider à améliorer les soins de santé et aider à une meilleure prise en charge des patients. Ces avancées sont dues à l'émergence des *transformers* dans les tâches de traitement du langage naturel, et aux percées qu'ils ont entraînées dans des domaines spécifiques comme le domaine médical. Tâches réalisées en exploitant des modèles pré-entraînés sur les données de santé. Nous verrons donc tout naturellement comment des modèles dérivés de BERT peuvent nous aider à envisager l'automatisation automatique d'IVAN, en les appliquant à la labellisation des articles qu'il doit indexer selon le thésaurus de l'ICM.

Il faut souligner ici une limitation de modèle issus de la famille de BERT quant à notre tâche d'indexation automatique d'IVAN. En effet, ces modèles ne prennent en entrée qu'un vecteur de 512 *tokens* maximum. C'est pourquoi, et même s'il existe des techniques pour contourner cette limitation, ces techniques dans la pratiques générant trop de bruits et entraînant une perte significative de cohérence de l'information, il faudrait se limiter à l'analyse des *abstracts* des articles seulement.

Or, jusqu'à maintenant, la base de connaissance IVAN n'a pas été conçu pour indexer les *abstracts*, il faudra la faire évoluer en ce sens s'il on envisage une indexation automatique. Il faut ajouter à cela, comme on le voit dans l'exemple commenté plus haut : si les documentalistes scientifiques essayent d'exploiter au maximum les résumés (*abstracts*) d'articles pour labéliser le document en vue de son indexation, cela n'est pas exclusif, il n'est pas rare qu'elles aient à affiner en balayant l'article dans son entièreté. C'est un des points de vigilance qu'il faut avoir à l'esprit et il faut être en mesure d'évaluer : si une indexation automatique est aussi qualitative et complète à partir des *abstracts* qu'une indexation à partir du corps de l'article. Evaluer si la perte d'information n'est pas trop importante et si elle peut

être compensée par le volume, l'effet de masse du traitement automatique de ces articles par l'intelligence artificielle.

Nous avons vu que les type de modèles comme BERT basés sur une architecture de *transformer* ont marqué une étape importante puisqu'ils ont considérablement augmenté notre capacité à faire de l'apprentissage par transfert en TAL.[50]

Pour rappel, l'apprentissage par transfert représente la capacité qu'un modèle neuronal entraîné sur une tâche (ex : prédiction de mot suivant) à pouvoir s'appliquer pour produire des résultats pertinents sur une autre tâche proche mais différente[51]. Par exemple, la prédiction d'une entité nommée c'est-à-dire un nom propre de personne, de lieu, d'organisation ou dans notre cas d'une molécule, d'une franchise.

Les algorithmes dérivés de BERT sont open source.

Un des premiers a été SciBERT. Il s'agit d'un modèle de représentation de langage entraîné sur des milliers d'articles scientifiques et développé par l'Allen Institute for Artificial Intelligence. Ce modèle est capable d'apprendre la structure d'un article scientifique, le contexte dans lequel des mots apparaissent et la sémantique des phrases.[52]

Sur ce principe la communauté de la recherche biomédicale a développé divers modèles dans leur domaine : comme BioBERT, BioELECTRA et BioALBERT.[53]

Chaque nouvelle architecture de modèle crée plus de possibilités d'ajustement pour un domaine spécifique. Par exemple, dans le domaine biomédical, des modèles inspirés du BERT tels que BioBERT, ClinicalBERT ou ELECTRA conduisent à BioELECTRA, etc. Les modèles spécifiques au domaine utilisent des ensembles de données spécifiques à leur domaine tels que PubMed, MIMIC III[54], etc. pour *finetuner* les modèles génériques ou les entraîner à partir de zéro.

3.2 QUEL MODELE CHOISIR ?

[50] https://www.linkedin.com/pulse/thoughts-nlp-biomedical-domain-akshay-chougule Consulté le 21/09/2022

[51] Amami, M., Mahé. Une approche fondée sur un algorithme open source.I2D.*op cit.*

[52] Ibid.

[53] Kalyan, K.S., Rajasekharan, A., Sangeetha S. AMMU : A Survey of Transformer-based Biomedical Pretrained Language Models. arxiv.org, 2021

[54] https://physionet.org/content/mimiciii/1.4/ Consulté le 18/10/2022

Si on souhaite résoudre un problème complexe de TAL biomédical, le fait d'avoir le bon point de départ peut non seulement augmenter la précision/les performances, mais peut également économiser des coûts et du temps de calcul importants. Le point de départ commun pourrait être de décider entre ces trois options :

1 - Choisir, la dernière architecture, l'entraîner à partir de zéro (par exemple PubMedBERT)

Ou bien,

2 - Choisir un modèle générique pré-entraîné (par exemple BERT) à *finetuner* sur notre ensemble de données biomédicales.

Ou bien,

3 - Choisir un modèle spécifique à un domaine pré-entraîné (par exemple BioBERT) pour le *finetuner* davantage sur un jeu de données biomédicales spécifiques à la tâche.[55]

En effet, les modèles BERT employés, en domaine spécialisé, semblent tous découler d'une stratégie assez simple : utiliser le modèle BERT originel comme initialisation puis poursuivre l'entraînement de celui-ci sur un corpus spécialisé. Il est clair que cette approche aboutit à des modèles plutôt performants. Cependant, il paraît raisonnable de penser qu'entraîner un modèle directement sur un corpus spécialisé, en employant un vocabulaire spécialisé, puisse aboutir à une meilleure pénétration du domaine, donc faire progresser les performances.

Mais, affiner ces modèles pour une tâche finale reste difficile, en particulier avec de petits ensembles de données labélisées, qui sont courants dans le TAL biomédical.

Ainsi, d'un côté des études systématiques sur le réglage fin *(finetuning)*, pour s'assurer de résultats stables ont été menés dans le TAL biomédical. Il a été montré que la performance du *finetuning* est sensible aux réglages utilisés lors de la phase de pré-entrainement, en particulier dans les domaines à faibles ressources.[56]

Enfin, d'autres études ont exploré de manière complète les techniques pour traiter l'instabilité du *finetuning* et il a été démontré que ces techniques améliorent de manière considérable les performances des processus de *finetuning* dans le TAL biomédical pour des applications ayant peu de ressources disponibles. Ces stratégies consistent à déterminer qu'elles

[55] https://www.linkedin.com/pulse/thoughts-nlp-biomedical-domain-akshay-chougule Consulté le 21/09/2022

[56] Tinn, R., Cheng, H., Gu, Y., Usuyama, N., *et al*. Fine-Tuning Large Neural Language Models for Biomedical Natural Language Processing. arxiv.org, 2021

sont les couches(*layers)* hérités du modèle pré-entrainé à garder tel quels et quels sont les paramètres à réentraîner[57].

Le point important à retenir est que, dans l'ensemble, le vocabulaire et le pré-entraînement sur des corpus spécifiques au domaine permettent de *finetuner* des modèles plus robustes. Sur la base de ces découvertes, un nouvel état de l'art a été établi sur un large éventail d'applications biomédical du champ du TAL.

4 BENCHMARK DE QUATRE METHODES ALTERNATIVES CONCERNANT LE PROBLEMATIQUE DE CLASSIFICATION MULTI-LABEL A PARTIR DE CORPUS MEDLINE

4.1 PUBMESH : UNE PREMIERE APPROCHE POUR AFFRONTER LA CLASSIFICATION MULTI-LABEL AVEC UN GRAND NOMBRE DE LABELS

Une équipe de chercheur de Stanford[58], a prouvé qu'utiliser un modèle d'une architecture proche celle de BERT était en mesure de rivaliser avec des curateurs humains, dans certains contextes, pour prédire les termes MeSH, nécessaires à l'indexation d'articles issues de PubMed.Leurs travaux sont intéressants à plusieurs titres.

D'une part le choix, de la *dataset.* Ils ont entraîné leur modèle sur une source de grande envergure : le *BioASQ Challenge Task* qui inclue 14 millions d'*abstracts,* labélisés en moyenne avec 14 termes MeSH chacun, pour un total de 29 351 termes uniques.

D'autre part, ils ont fait le choix d'utiliser des modèles proches de BERT : c'est-à-dire qu'ils ont utilisés quatre modèles (qu'ils ont fusionnés) avec une approche bidirectionnelle donc en mesure de saisir le contexte global d'un mot. De plus dans leur architecture, ils ont utilisé un mécanisme d'attention pour dégager des relations et des *patterns* linguistiques, donc une représentation de haut niveau des dépendances des concepts clés identifiés et de leur contexte. Enfin, ils ont utilisé des *Word embeddings* pré-entrainés sur ce corpus.

On voit donc, qu'ils ont été en mesure de tirer bénéfice des principaux atouts que nous livrerait une architecture *transformer* et un modèle pré-entraîné comme BERT, sans toutefois obtenir des résultats à l'état de l'art qu'offre maintenant un tel modèle. C'est la limite de cette article, il se base sur des outils qui sont maintenant largement dépassés par cette nouvelle génération de modèles basés sur les *transformers* et pré-entraînés, même s'il en conserve les intuitions. C'est malgré tout un bon point de départ, qui démontre que notre approche est non seulement justifiée mais également performante si on la compare à la curation humaine.

Dans leurs résultats, ils présentent plusieurs exemples de cas où leur modèle à livrer un meilleur travail qu'un curateur humain mais également des exemples où ça été l'inverse, en

[57] Ibid.

[58] Thomas, K., Paul, R., Kanzawa, M. PubMeSH : Extreme Multi-label Classification of Biomedical Research. Standford, 2018

expliquant pourquoi, ce qui donne un aperçu éclairant du potentiel et des limites d'une curation automatisé par l'intelligence artificielle.
Pour illustrer nous proposons deux exemples tirés de leur article.

- Le premier où le modèle a prédit des *labels* pertinents qu'un curateur humain a omis :

Model: [Fracture Fixation, Osteoporosis] **Expert:** [Lumbar Vertebrae, Pain Measurement].

Or d'après les auteurs la réparation de la fracture et l'ostéoporose sont tous les deux mentionnés dans l'*abstract*. Ce sont des *labels* plus précis que « vertèbre lombaire » et ils sont à la source de la douleurs que les auteurs de l'article de PubMed cherchent à mesurer.

- Le deuxième, a contrario est un cas où le modèle a omis des *labels* importants :

Model: [Brazil, Food, Fruit] **Expert:** [Brazil, Fruit, Meat, Vegetables, Energy Intake, Eating, Feeding Behavior, Obesity, Body Mass Index]

Le modèle a correctement identifié les *labels* « Brésil » et « fruit », et le sujet générique : « nourriture » qui est manquant dans les *labels* du curateur humain. Cependant, il a échoué à identifier les aspects cliniques de l'étude : il ne prédit pas : l'« obésité », l'« index de masse corporel », ou la « consommation d'énergie ». Les chercheurs supposent que les W*ord embeddings* pour la « nourriture » sont trop présent dans l'abstracts, et qu'ils sont trop similaires aux *Word embeddings* des termes MeSH cliniques qui étaient à identifier dans l'*abstract*, ce qui les auraient rendus opaques au mécanisme d'attention quand il a été appliqué.

4.2 ML-NET : UNE ARCHITECTURE COMPLEXE ALTERNATIVE FACE AU GRAND NOMBRE DE LABELS

Une autre équipe de chercheurs, venant de grandes institutions de la santé américaine et associé au département de l'informatique biomédicale de l'université du Texas, a proposé ML-Net, un nouveau cadre d'apprentissage profond, pour la classification *multi-label* de textes biomédicaux. ML-Net combine un réseau de prédiction de *labels* avec un mécanisme automatisé de prédiction du nombre de *labels* pour produire un ensemble optimal de *labels*, en exploitant à la fois le score de confiance prédit de chaque *label* et les informations contextuelles du document cible. C'est cette combinaison qui fait son originalité. Contrairement aux méthodes d'apprentissage automatique traditionnelles, ML-Net ne nécessite pas d'efforts humains dans l'ingénierie des fonctionnalités (*feature engeenering*). C'est-à-dire qu'elle n'a pas besoin d'avoir à faire à un humain pour sélectionner les informations préalables nécessaire à ses prédictions. Plus trivialement, on peut dire qu'elle n'a pas besoin d'avoir recours à un humain pour « bidouiller » les *datasets* afin d'extraire ce qui est pertinent, ou « coder en dure » (*hardcoding)* arbitrairement les paramètres pour « forcer » le résultat, ces derniers sont appris grâce à l'algorithme d'apprentissage profond. ML-Net constitue une approche très efficace et évolutive pour les tâches avec un grand nombre de *labels* : pas besoin de créer des classifieurs individuels pour chaque *label* distinct ce qui était l'approche traditionnelle. Dans la

classification de texte *multi-label*, chaque document textuel se voit attribuer un ou plusieurs *label*. Ainsi auparavant pour résoudre le problème de la classification *multi-label*, on décomposait le problème en une multitude de tâche de classification binaire : pour chaque *label* le modèle devait décider si le *label* était pertinent ou non et par conséquent chaque *label* était indépendant des autres et aucune corrélations potentielles, entre ces *labels*,apprises. ML-NET est lui capable d'estimer dynamiquement le nombre de *labels* en fonction du contexte du document de manière plus systématique et précise. Ceci est accompli en tirant parti à la fois du score de confiance prévu de chaque *label* et des informations contextuelles approfondies dans le document cible. Ces informations contextuelles sont les *Word embedding* pré-entrainés à partir du modèle ELMo (un modèle antérieur à BERT) et qui sont *finetunés* sur leurs *datasets*. Là aussi, un mécanisme d'attention a été utilisé pour dégager les représentations des documents utilisés comme *inputs* de leur modèle.

Ils ont évalué ML-Net sur trois corpus indépendants et accessibles au public dans deux types de genre de texte : la littérature biomédicale et les notes cliniques. ML-Net est comparé à plusieurs modèles de base d'apprentissage automatique concurrents. Les résultats d'analyse comparatifs montrent que ML-Net se compare favorablement aux méthodes en pointe en matière de classification *multi-label* de textes biomédicaux. ML-NET s'avère également robuste lorsqu'il est évalué sur différents genres de texte en biomédecine. En tant que tâche importante qui a de larges applications en biomédecine, un certain nombre de méthodes de calcul différentes ont été proposées. Beaucoup de ces méthodes, cependant, n'ont qu'une précision ou une efficacité modeste et un succès limité dans l'utilisation pratique.

Si ce nouveau cadre d'apprentissage profond ML-Net, pour la classification *multi-label* de textes biomédicaux, ne tire pas profit des nouvelles avancées apportées par les *transformers*, ces derniers sont appelés de leurs vœux comme pistes d'amélioration pour de nouvelles recherches, par les auteurs à la fin de l'article. [59]

4.3 BERTMESH : UN MODELE QUI NE SE LIMITE PAS QU'AUX ABSTRACTS

En 2020, des universitaires chinois ont utilisé pour la première un modèle de *transformer* pré-entraîné pour résoudre le problème de l'indexation automatique de termes MeSH. Un des intérêt de leur expérimentation a été de tirer profit de longues parties des articles de PubMed. Ainsi, en plus, du titre et de l'*abstract*, comme c'était l'usage, ils se sont servi des quatre sections les plus longues parmi ces cinq sections standards d'un article qui sont : « Introduction, Methods and Materials, Result and Experiment, Conclusion and Summary ». Dans leur étude, ils se servent de BioBERT pour dégager les *Word embeddings*. Ce modèle utilise les paramètres initiaux de BERT puis le *finetune* sur les citations MEDLINE comprenant le titre et l'*abstract* et sur les articles complets de PubMed Central. Leur modèle complet BERTMeSH utilise également un mécanisme d'attention appliqué pour se concentrer sur les portions de textes les plus significatives par rapport à un *label* MeSH particulier pour un article donné.

[59] Du, J., Chen, Q., Peng, Y., Xiang, Y., *et al*. ML-Net: multi-label classification of biomedical texts with deep neural networks. Journal of the American Medical Informatics Association, Vol. 26, n° 11, 2019. p. 1279–1285

BERTMeSH a eu des résultats, statistiquement signifiants, qui ont montré qu'adopter un modèle de *transformer* pour l'indexation automatique des articles PubMed surpassait les capacités des modèles faisant alors références. Mais surtout, ils ont pointé que le modèle était d'autant plus puissant qu'il utilisait les ressources des articles complets de PubMed Central. Enfin, le modèle a fait preuve de robustesse, même entraîné sur des articles avec des sections manquantes. Cerise sur le gâteau, le modèle n'a besoin que de cinq minutes pour prédire les étiquettes de 20 000 articles !

4.4 PUBMEDBERT : LE MODELE LE PLUS EFFICIENT POUR NOTRE PROBLEMATIQUE

Dernier point et un des plus important, une équipe de chercheurs appartenant à la *National Library of Medicine,* a pour la première fois, utilisé en 2021 un modèle de *transformer spécifique au domaine biomédical* pour une tâche de *text ranking* dans le cadre de l'indexation automatique des articles PubMed[60]. Ce n'est pas directement une tâche de classification *multi-label*, mais une tâche en aval, ici, de la classification *multi-label* d'un réseau de neurone plus classique (un CNN ou *Convolutional Neural Network*[61]) qui consiste à classer ses prédictions par ordre de probabilité (« *ranker* ») en ne gardant que les plus pertinentes. Et donc d'en renforcer l'efficience. C'est une tâche nécessaire étant donné le grand nombre de termes MeSH et la présence de nombreux termes peu fréquents. Jusqu'à maintenant les systèmes d'automatisation de cette tâche étaient peu performants et nécessitaient beaucoup d'opérations faites à la main *(features engineering)*.

Mais surtout, le nouveau modèle qu'ils utilisent : PubMedBERT, à la différence du modèle que nous avons vu précédemment : BioBERT, a, non seulement, été entraîné sur les *abstracts* de PubMed et les articles complets disponibles sur PubMed Central, mais, c'est là où il est unique, il n'a pas seulement été *finetuné* sur ce corpus, il a été entraîné à partir de zéro et a donc acquis le vocabulaire spécialisé du domaine biomédical et non les 30 000 tokens du domaine général comme c'est le cas pour BioBERT qui en a hérité de BERT. On peut donc affirmer, avec les auteurs, que PubMedBERT est actuellement le modèle le plus performant pour les corpus biomédicaux.

L'étude a montré que PubMedBERT était un bon choix de modèle parce qu'il a prouvé qu'il était le meilleur modèle pour encoder efficacement l'information biomédical dans ses

[60] Rae, A.R., Mork, J. G., Demner-Fushman, D. A Neural Text Ranking Approach for Automatic MeSH Indexing.National Library of Medicine, 2021

[61] Un réseau neuronal convolutif est un type de réseau de neurones artificiels dans lequel le motif de connexion entre les neurones est inspiré par le cortex visuel des animaux. L'*input* est « pavé » c'est-à-dire découpé en petites zones, qui seront traitées individuellement par un neurone artificiel (qui effectue une opération de filtrage en associant un poids à chaque élément de l'*input*).

Word embeddings. Ce choix a été évalué, lors du *challenge* BioASQ[62] de 2021, sur la tâche 9a[63] qui est la référence pour comparer les modèles s'attaquant au problème de l'indexation automatique des articles PubMed.

Nous avons introduit préalablement le modèle PubMedBERT. L'équipe de chercheurs qui l'a mis au point, une équipe de Microsoft, dans une étude approfondi[64], a montré que le pré-entraînement de grands modèles de langage neuronaux, tels que BERT, a conduit à des gains impressionnants sur de nombreuses tâches du traitement du langage naturel (TAL). Cependant, la plupart des efforts de pré-entraînement se concentrent sur des corpus de domaine général, où prédominent les flux d'actualités et le Web.

L'hypothèse de départ, généralement admise est que même le pré-entraînement sur un domaine spécifique peut bénéficier de modèles de langage du domaine général.

Dans cet article, l'équipe de chercheur remet en cause cette hypothèse en montrant que pour les domaines contenant beaucoup de texte non labélisés, comme la biomédecine, le pré-entraînement des modèles de langage à partir de zéro entraîne des gains substantiels par rapport au pré-entraînement continue à partir du langage du domaine général des modèles.

Pour leur enquête, cette équipe a compilé un ensemble de référence de données biomédicales complet en TAL à partir de données qui étaient jusqu'à lors dispersées en plusieurs corpus de références différents. Leurs expériences montrent que la pré-entraînement spécifique à un domaine constitue une base solide pour un large éventail de domaines biomédicaux. Ils ont abouti à de nouveaux résultats de pointe, pour de nombreuses tâches de TAL dans tous les domaines.

Le schéma, ci-dessous[65], illustre les stratégies d'entrainement entre des modèles BERT qui ont été entrainés sur des données textuelles du domaine général en l'occurrence Wikipedia, Bookcorpus, et des articles de presse accessibles en ligne : en bleu. Puis, sur des données spécifiquement biomédicales : en orange. Enfin, des modèles qui ont été entraînés de A à Z sur des données biomédicales, ici en l'occurrence, issues de PubMed. Dans leur article, les auteurs

[62] BioASQ est une compétition qui porte sur l'indexation sémantique biomédicale et les systèmes de questions-réponses (question answering ou QA) à une large échelle.

[63] La tâche 9a de BioASQ consiste à classifier les nouveaux documents PubMed, écrits en anglais, avant qu'un curateur PubMed ne les classent manuellement. Les classes sont les *Headings* du sujet qui sont actuellement utilisés pour indexer les *abstracts* manuellement. Les modèles participant ont la possibilité d'être entraînés sur tout l'historique des *abstracts,* annotés manuellement, de PubMed. C'est la confrontation des annotations manuelles et automatiques pour un même *abstract* qui permet d'évaluer la performance d'un modèle par rapport à la curation humaine.

[64] Tinn, R., Cheng, H., Gu, Y., Usuyama, N., *Et al*. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing.Arxiv.org, 2020

[65] Ibid.

visent à démontrer l'impact de ces stratégies sur le vocabulaire final du modèle et sur la qualité des *Word embeddings* qui en découle.

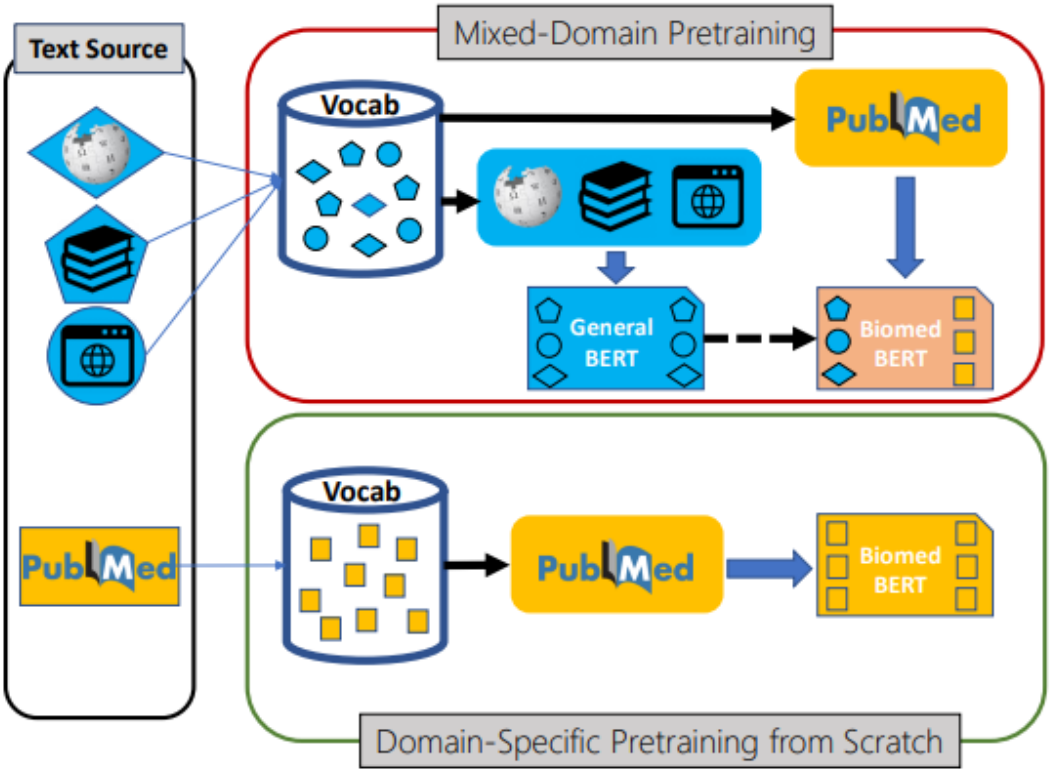


Figure 27 : Représentation des divergences du vocabulaire des modèles BERT selon leurs corpus sources

Pour comprendre cet impact il faut aborder le problème de la tokenisation lors de la phase de pré-traitement (*preprocessing).* Les *tokens* sont obtenus par un algorithme de segmentation de sous-mots (les *tokens*) utilisé dans le traitement du langage naturel qui dans le cadre de BERT s'appelle WordPiece. Le vocabulaire est initialisé avec des caractères individuels dans la langue, puis les combinaisons de symboles les plus fréquentes dans le vocabulaire sont ajoutées de manière itérative au vocabulaire. Il est important de noter à ce stade que linguistiquement parlant ces sous-mots ne sont pas des morphèmes, car ils ne portent pas de sens en soi, ce sont des signifiants vides de sens. En effet, l'algorithme commence par calculer la fréquence d'une paire de *tokens* en commençant par les caractères qui sont fusionnés à chaque itération par ordre de fréquence jusqu'à reconstituer un mot. Le vocabulaire comprend donc tous les caractères de la langues anglaise, les mots les plus fréquent du corpus d'entraînement, puis les morceaux de mots : les pairs ou les ensembles de pairs les plus courant pour former la totalité des mots inconnus (c'est-à-dire qui ne sont pas compris dans le vocabulaire qui est, je le rappelle, de taille préfixée). La limite étant que pour ces mots non compris dans le vocabulaire, qui vont être reconstruits par une combinaison de N *tokens*, la génération d'un

Word embedding propre n'est pas possible. Ils auront à la place : un *Word embedding* qui résulte de la combinaison des *Word embeddings* des *tokens* qui le constituent. Par conséquent, les représentations pour les mots spécifiques au domaine cible, ici le biomédical, auront été dégagés de contextes souvent très éloignés et non corrélés. Ce qui conduit à une perte, inévitable, de sens intrinsèque, substantielle. Si on prend pour exemple, dans le tableau ci-dessous[66], le mot « cardiomyocyte », on observe que BERT le décompose en cinq *tokens* : « *card* », « *iom* », « *yo*, » « *cy* », « *te* ». Pour sa part, SciBERT en deux *tokens* : « *cardiomy-ocyte* ». Enfin, PubMedBERT a inclus dans son vocabulaire « *cardiomyocyte* » en un seul *token*.

Biomedical Term	Category	BERT	SciBERT	PubMedBERT (Ours)
diabetes	disease	✓	✓	✓
leukemia	disease	✓	✓	✓
lithium	drug	✓	✓	✓
insulin	drug	✓	✓	✓
DNA	gene	✓	✓	✓
promoter	gene	✓	✓	✓
hypertension	disease	hyper-tension	✓	✓
nephropathy	disease	ne-ph-rop-athy	✓	✓
lymphoma	disease	l-ym-ph-oma	✓	✓
lidocaine	drug	lid-oca-ine]	✓	✓
oropharyngeal	organ	oro-pha-ryn-ge-al	or-opharyngeal	✓
cardiomyocyte	cell	card-iom-yo-cy-te	cardiomy-ocyte	✓
chloramphenicol	drug	ch-lor-amp-hen-ico-l	chlor-amp-hen-icol	✓
RecA	gene	Rec-A	Rec-A	✓
acetyltransferase	gene	ace-ty-lt-ran-sf-eras-e	acetyl-transferase	✓
clonidine	drug	cl-oni-dine	clon-idine	✓
naloxone	drug	na-lo-xon-e	nal-oxo-ne	✓

Figure 28 : Comparaison de la complétude des vocabulaires de plusieurs modèles BERT

Cela a pour conséquence, ni plus ni moins, la résolution ou non de la tâche finale. Dans les deux exemples, ci-dessous, qui évaluent le modèle BioBERT et le modèle PubMedBERT sur deux tâches de TAL biomédical différentes, la tokenisation du mot « *epithelial* », « *serine* » et celle d'« *agonistic* » diffère sensiblement. BioBERT dilue le sens des mots décisifs pour interpréter la phrase correctement, car elle a hérité les *tokens* du domaine général. Au contraire, le vocabulaire spécifiquement biomédical de PubMedBERT lui permet d'apprendre des *patterns* grâce à un mécanisme d'attention. Il est représenté dans ce graphique par les traits en violet et le marquage en bleu qui sont d'intensité plus ou moins faible pour symboliser les liens

[66] Ibid.

tissés entre les *tokens*. Ce mécanisme d'attention est de bien meilleure qualité dans son cas et lui permet, du coup, de faire des prédictions valides.

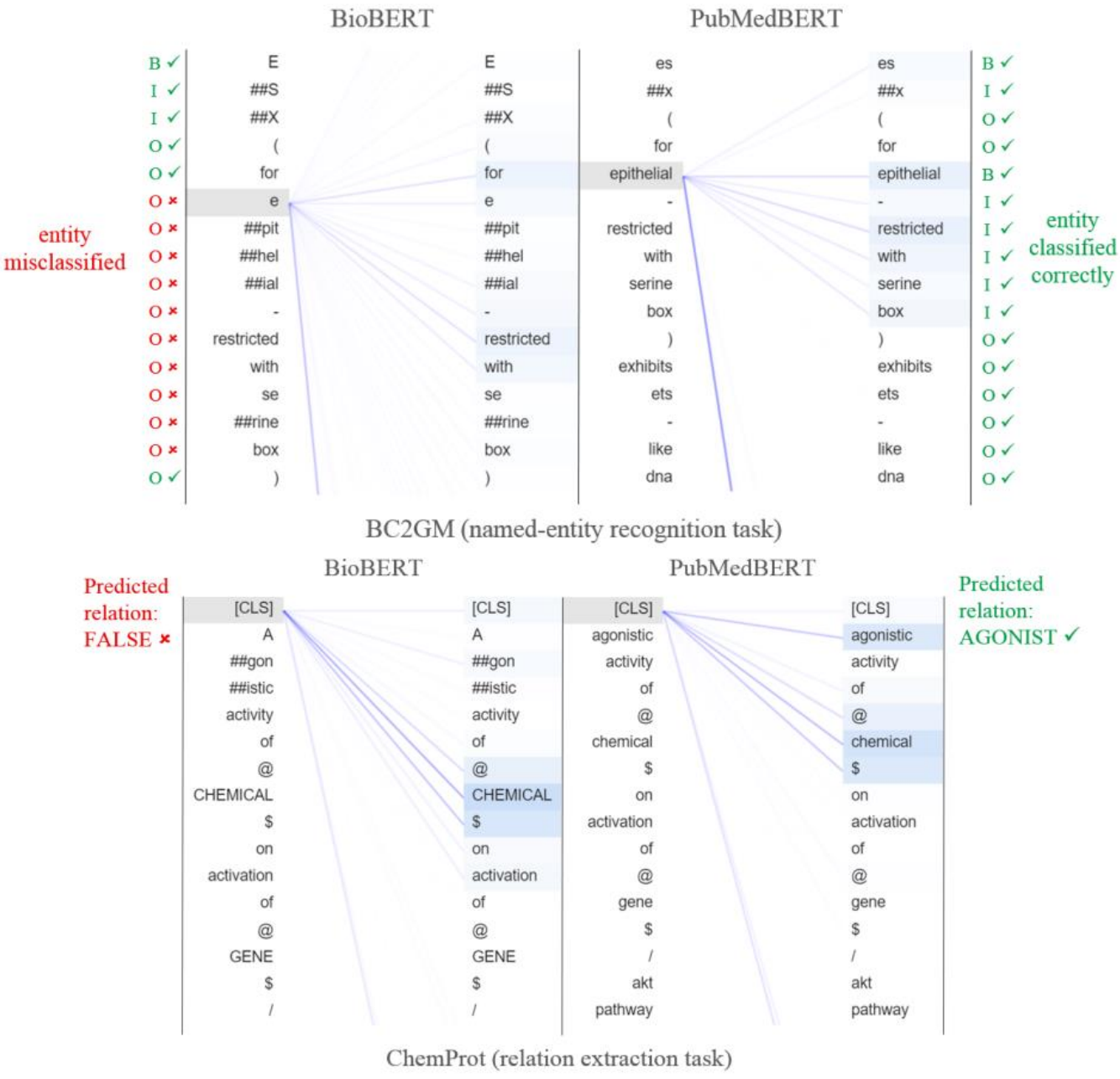


Figure 29 : Comparaison de la dernière couche des modèles BioBERT et PubmedBERT

Dans les graphiques du haut[67], les deux modèles ont été évalués sur deux tâches que nous étudieront, plus en détail, plus loin dans le mémoire. Pour les définir brièvement : la première est une tâche de reconnaissance d'entité nommée. En partant d'une phrase donnée, la tâche de reconnaissance d'entité nommée, a pour but d'identifier les entités nous intéressant. C'est une tâche très similaire à la classification, mais là, à chaque *token* on attribue une *label* pour déterminer, si oui ou non, il est assimilable à une entité donnée. Ici les lettre B,I,O classifie chaque *token* comme le début d'une entité (B), l'intérieur d'une entité(I), ou son extérieur(O). Là tâche d'extraction de relations consiste à extraire des relations sémantiques d'un texte.

On a plus particulièrement sur ces deux vues superposés, la représentation graphique d'une des dernières couches du modèle où se détecte les *patterns* sémantiques de haut niveau. Surlignées en bleu, les corrélations les plus significatives (je le redis : plus la couleur bleue est intense, plus les mots sont corrélés) et les traits violets symbolisent les relations, les dépendances relatives d'un *token* avec le reste de la phrase. Plus la couleur violette est forte, plus les termes sont liés.

La phrase est en miroir ici pour illustrer le mécanisme d'attention d'un mot par rapport au reste de la phrase (y compris lui-même). Pour être plus précis on calcul la « similarité » d'un *token* par rapport aux autres *tokens* de la phrase pour déterminer son contexte. Plus le score de probabilité ou d'attention est haut plus les *tokens* sont corrélés. Le *token* CLS dans le dernier exemple est un *token* spécial qui désigne le début de la phrase. Calculer un score d'attention, des autres *tokens*, par rapport à lui-même, permet de dégager les éléments signifiant de la phrase pour prédire la relation « *agonist* » ou en français « agoniste » (c'est-à-dire qui a un effet identique) entre une substance chimique et un gène. Dans cet exemple, encore une fois, BioBERT a raté l'information importante qui était l'adjectif « *agonistic* » qu'il n'avait pas dans son vocabulaire sous forme entière mais sous la forme de trois *tokens* différents (« *A* », « *##gone*[68],*##istic*) dépourvus de sens. Ces deux exemples démontre à quel point il est crucial d'avoir un vocabulaire entraîné de zéro sur un corpus appartenant au domaine où on veut appliquer le modèle.

Les auteurs de l'étude sont aussi à l'origine d'un benchmark (BLURB ou Biomedical Language Understanding & Reasoning Benchmark) qui se veut complet en réunissant les principales *datasets* de référence dans le petit monde du TAL biomédical anglosaxon et les tâches qui leurs sont associées. Il est bon de noter que la quasi-totalité de ces *datasets* ont pour source PubMed. Et les tâches concernés sont aussi divers que la reconnaissance d'entités nommées, l'extraction de relations, la classification de documents, et les questions-réponses (QA). Pour accélérer la recherche, ils ont créé, à partir de là, un *leaderboard*, afin d'évaluer de manière rigoureuse et systématique les modèles s'attaquant au TAL biomédical.

[67] Ibid.

[68] « ## » ici est utilisé pour signifier que *le token* est un suffixe.

Une étude commis par une équipe de chercheurs appartenant à Facebook[69], nuance l'étude précédente, elle montre qu'un modèle de la même famille plus robuste que BERT comme RoBERTa, développé par Facebook, peut être suffisant pour un grand nombre de tâches de TAL biomédical. Ce qui a son importance pour notre cas d'étude, comme nous le verrons plus loin dans le mémoire. Même si, elle concède que l'utilisation d'un vocabulaire entraîné sur le domaine spécialisé qu'est le biomédical fait la différence. Elle a produit cette analyse à partir d'un *benchmark* de 18 tâches qui, à la différence de BLURB, incluent des *datasets* de notes cliniques. Les auteurs, en ouverture, proposent d'évaluer leurs modèles sur BLURB.

4.5 BILAN D'ETAPE

4.5.1 COMMENT EVALUER UN MODELE ?

On peut tenter d'établir un premier bilan avec les évaluations des modèles dérivés de BERT récapitulés dans, le graphique et les deux tableaux à la fin de cette partie.

Mais d'abord il faut expliciter ce que veut dire évaluer en *Machine Learning.* Pour évaluer la performance des modèles, il faut une métrique permettant de les comparer. Le *F1 score* est la plus utilisé et nous allons la détailler ci-dessous.

Le *F1 score* tient compte, non seulement du nombre d'erreurs de prédiction commises par le modèle, mais examine également le type d'erreurs commises.

- La *Precision* est le premier constituant du score F1. Voici la formule qui permet sont calcul :

$$Precision = \frac{\#\ of\ True\ Positives}{\#\ of\ True\ Positives + \#\ of\ False\ Positives}$$

[69] Lewis, P.S.H., Ott, M., Du, J., *et al*. Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. openReview.net,2020

On peut interpréter cette formule de la manière suivante. Dans tout ce qui a été prédit comme positif, la *Precision* compte en pourcentage le nombre de ceux qui sont correct : un modèle imprécis peut trouver beaucoup de positifs, mais sa méthode de sélection comporte beaucoup de bruit, c'est-à-dire qu'i il détecte également à tort de nombreux positifs qui ne sont pas réellement positifs.
Un modèle précis est très « pur » : peut-être ne trouve-t-il pas tous les positifs, mais ceux que le modèle classe comme positifs sont très probablement corrects.

Le *Recall* est le second constituant du score F1. Voici la formule :

$$Recall = \frac{\#\ of\ True\ Positives}{\#\ of\ True\ Positives + \#\ of\ False\ Negatives}$$

On peut interpréter cette formule de la manière suivante. Dans tout ce qui est réellement positif, combien le modèle a-t-il réussi à en trouver ?

Un modèle avec un *Recall* élevé réussit bien à trouver tous les cas positifs dans les données, même s'il peut également identifier à tort certains cas négatifs comme des cas positifs. Un modèle à faible *Recall* n'est pas en mesure de trouver tous (ou une grande partie) des cas positifs dans les données.

Precision et Recall sont les deux éléments constitutifs du *F1 score*. L'objectif du score F1 est de combiner les mesures *de Precision* et de *Recall* en une seule mesure. Le *F1 score* a été conçu pour évaluer des données qui ont la problématique d'être déséquilibrées, ce qui est notre cas. Le compromis *Precision-Recall* représente le fait que dans de nombreux cas, on peut modifier un modèle pour augmenter la *Precision* au prix d'un *Recall* inférieur, ou d'autre part augmenter le rappel au prix d'une *Precision* moindre. Le *F1 score* est défini comme la moyenne harmonique de la *Precision* et du *Recall*. Pour rappel, la moyenne harmonique est une métrique alternative à la moyenne arithmétique plus courante. Il est souvent utile lors du calcul d'un taux moyen. La formule pour le *F1 score* est la suivante :

$$F1\ score = 2 * \frac{Precision\ * Recall}{Precision + Recall}$$

NB : D'autres variantes existent comme le *Micro F1 score* et le *Macro F1 score*. Une simple mesure d'*Accuracy* est aussi utilisé pour évaluer certaines tâches[70].

Pour une vue d'ensemble des mesures agrégées pour BLURB, dans les tableaux comparatifs que nous analysons dans ce bilan, il faut se référer à ce tableau ci-dessous[71] :

Dataset	Task	Train	Dev	Test	Evaluation Metrics
BC5-chem	NER	5203	5347	5385	F1 entity-level
BC5-disease	NER	4182	4244	4424	F1 entity-level
NCBI-disease	NER	5134	787	960	F1 entity-level
BC2GM	NER	15197	3061	6325	F1 entity-level
JNLPBA	NER	46750	4551	8662	F1 entity-level
EBM PICO	PICO	339167	85321	16364	Macro F1 word-level
ChemProt	Relation Extraction	18035	11268	15745	Micro F1
DDI	Relation Extraction	25296	2496	5716	Micro F1
GAD	Relation Extraction	4261	535	534	Micro F1
BIOSSES	Sentence Similarity	64	16	20	Pearson
HoC	Document Classification	1295	186	371	Micro F1
PubMedQA	Question Answering	450	50	500	Accuracy
BioASQ	Question Answering	670	75	140	Accuracy

Table 3. Datasets used in the BLURB biomedical NLP benchmark. We list the numbers of instances in train, dev, and test (e.g., entity mentions in NER and PICO elements in evidence-based medical information extraction).

Tableau 1 : LES DIFFERENTS TYPES DE MESURE UTILISEES POUR BLURB

4.5.2 ML-NET VS DES MODELES DERIVES DE BERT

Au préalable, voyons le différentiel entre un modèle connexionniste, non basé sur l'architecture *transformer* et des modèles BERT. Ici nous prendrons le cas d'étude de modèles appliqués à un sous-produit de PubMed, la base LitCovid consacrée aux articles PubMed

[70] Pour une pour plus de détail sur la différence entre ces mesures nous vous renvoyons à ce très pédagogique article de blog :

https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f, Consulté le 16/11/2022

Et pour plus d'exotisme, même ci-cela est plus anecdotique, Voir cette vidéo : https://www.youtube.com/watch?v=ciqcsWqIHIQ, Consulté le 16/11/2022

[71] Tinn, R., Cheng, H., Gu Y. Usuyama N., *et al*. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. Arxiv.org, 2020

exclusivement constitués d'articles sur la Covid. Dans cette vue à l'échelle[72], on peut observer l'écart entre ML-Net en rose : un modèle performant, que nous avons présenté, mais non dérivé de BERT et des modèles dérivés de BERT.

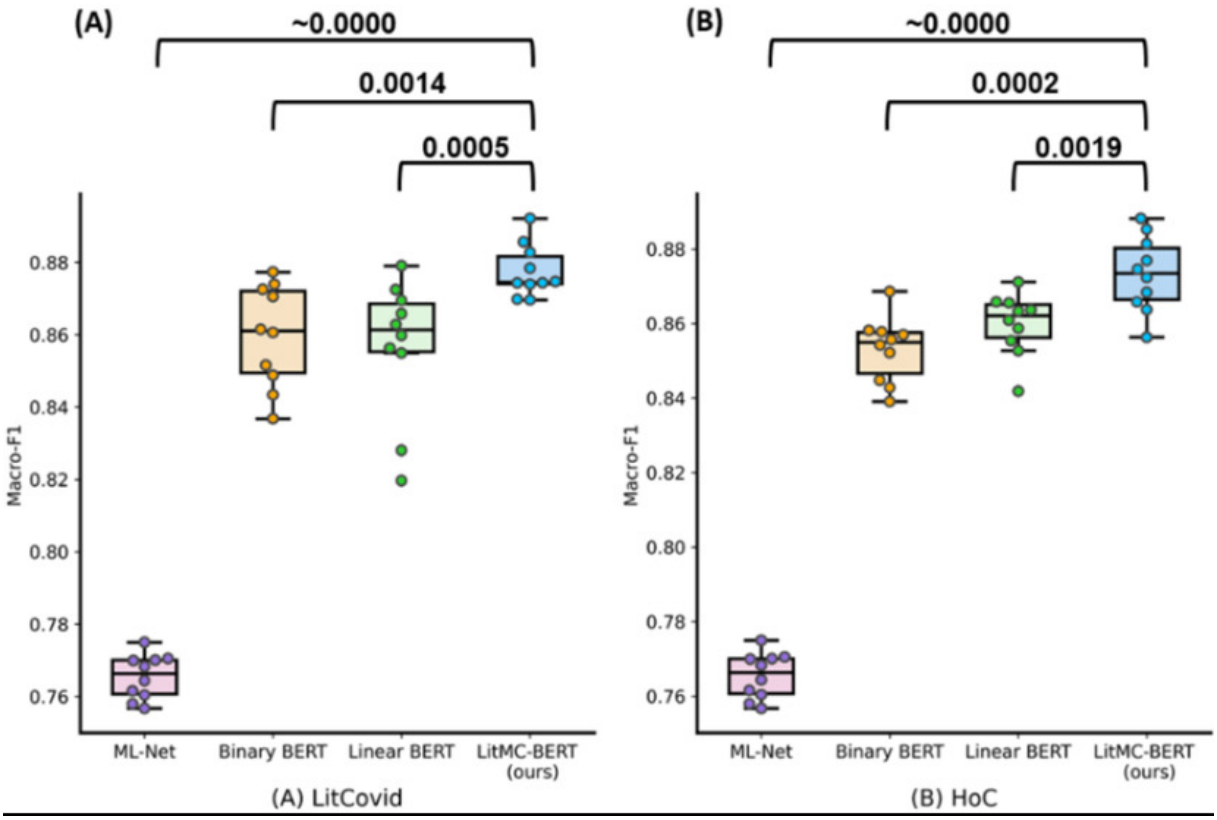


Figure 30 : Comparaison macro F1 score ML-Net et d'autres modèles dérivés de BERT

4.5.3 PREMIER TABLEAU COMPARATIF : LES MODELES BASES SUR DES DOMAINES GENERIQUES VS CEUX BASES SUR DES DOMAINES SPECIFIQUES[73]

[72] Chen, Q., Du, J., Allot, A., Lu, Z. LitMC-BERT: transformer-based multi-label classification of biomedical literature with an application on COVID-19 literature curation. IEEE/ACM Trans Comput Biol Bioinform, 2022

[73] Tinn, R., Cheng, H., Gu Y. Usuyama N., *et al*. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. Arxiv.org, 2020

	BERT		RoBERTa	BioBERT	SciBERT		ClinicalBERT	BlueBERT	PubMedBERT
	uncased	cased	cased	cased	uncased	cased	cased	cased	uncased
BC5-chem	89.25	89.99	89.43	92.85	92.49	92.51	90.80	91.19	**93.33**
BC5-disease	81.44	79.92	80.65	84.70	84.54	84.70	83.04	83.69	**85.62**
NCBI-disease	85.67	85.87	86.62	**89.13**	88.10	88.25	86.32	88.04	87.82
BC2GM	80.90	81.23	80.90	83.82	83.36	83.36	81.71	81.87	**84.52**
JNLPBA	77.69	77.51	77.86	78.55	78.68	78.51	78.07	77.71	**79.10**
EBM PICO	72.34	71.70	73.02	73.18	73.12	73.06	72.06	72.54	**73.38**
ChemProt	71.86	71.54	72.98	76.14	75.24	75.00	72.04	71.46	**77.24**
DDI	80.04	79.34	79.52	80.88	81.06	81.22	78.20	77.78	**82.36**
GAD	80.41	79.61	80.63	82.36	82.38	81.34	80.48	79.15	**83.96**
BIOSSES	82.68	81.40	81.25	89.52	86.25	87.15	91.23	85.38	**92.30**
HoC	80.20	80.12	79.66	81.54	80.66	81.16	80.74	80.48	**82.32**
PubMedQA	51.62	49.96	52.84	**60.24**	57.38	51.40	49.08	48.44	55.84
BioASQ	70.36	74.44	75.20	84.14	78.86	74.22	68.50	68.71	**87.56**
BLURB score	76.11	75.86	76.46	80.34	78.86	78.14	77.29	76.27	**81.16**

Table 6. Comparison of pretrained language models on the BLURB biomedical NLP benchmark. The standard task-specific models are used in the same fine-tuning process for all BERT models. The BLURB score is the macro average of average test results for each of the six tasks (NER, PICO, relation extraction, sentence similarity, document classification, question answering). See Table 3 for the evaluation metric used in each task.

Tableau 2 : TABLEAU COMPARATIF DE DIFFERENT MODELES BERT SUR BLURB

Première remarque concernant ce premier tableau : on observe que les modèles dérivés de BERT ont tous d'excellent résultats. Les plus faibles ont des résultats égaux ou supérieurs à 76 % de réussite et on observe un écart de 6% entre le moins bon élève et le meilleur élève. Avec un minimum de 76,16 % pour le modèle de base originale de BERT et un maximum de 81,16% pour le modèle PubMedBERT de *base,* si on compare leur moyenne générale.

Par sa stratégie de pré-entraînement à partir de zéro sur le domaine spécifiquement biomédical, PubMedBERT surpasse systématiquement tous les autres modèles BERT dans la plupart des tâches de TAL biomédicales, souvent avec une marge significative. Les gains sont les plus importants par rapport aux modèles BERT pré-entraînés à l'aide de texte hors domaine. Notamment, alors que le corpus de pré-entraînement est le plus important pour RoBERTa, ses performances sur les tâches de TAL biomédicales sont parmi les pires. Elles sont similaires au modèle BERT original.

Les modèles utilisant du texte biomédical pendant leur phase de pré-entraînement fonctionnent généralement mieux. Cependant, mélanger des données hors domaine pendant la phase de pré-entraînement conduit généralement à de moins bonnes performances. En particulier, même si les notes cliniques sont plus pertinentes, car elles se rapprochent du domaine biomédical que les textes du domaine général, les ajouter ne confère aucun avantage, comme en témoigne les résultats de ClinicalBERT et BlueBERT.

Sans surprise, BioBERT est le plus proche de PubMedBERT, car il utilise les textes de PubMed pour son pré-entraînement. Cependant, en procédant à partir de zéro à un pré-entraînement spécifique au domaine, ce qui lui permet de capturer le vocabulaire biomédical issu de PubMed, PubMedBERT est capable d'obtenir des gains constants par rapport à BioBERT dans la plupart des tâches.

Si on étudie les performances des modèles *LARGE* par rapport aux modèles de *BASE*, on constate une contreperformance de la version *LARGE* de BioBERT par rapport à sa version de *BASE*. Mais, surtout, que les modèles PubMed *LARGE* et PubMedELECTRA *LARGE*, qui eux aussi bénéficient d'un pré-entraînement sur des données biomédicales obtiennent des gains de performances significatifs par rapport à leurs version de *BASE*.

4.5.4 DEUXIEME TABLEAU COMPARATIF : LES MODELES DANS LEURS VERSIONS LARGES VS LEURS VERSIONS DE BASE[74]

	BioBERT -LARGE	BlueBERT -LARGE	PubMedBERT -LARGE	PubMedELECTRA -LARGE
BC5-chem	93.05	90.24	**93.23**	92.90
BC5-disease	84.97	82.93	**85.77**	84.82
NCBI-disease	**88.76**	86.44	88.25	87.93
BC2GM	84.21	80.86	**84.72**	83.87
JNLPBA	78.83	77.59	**79.44**	78.77
EBM PICO	73.81	72.43	73.61	**73.95**
ChemProt	77.79	71.31	**78.77**	76.80
DDI	81.53	78.99	**82.39**	78.92
GAD	82.47	75.80	83.57	**83.93**
BIOSSES	91.53	86.18	**92.73**	90.33
HoC	81.57	81.35	**82.57**	82.37
PubMedQA	55.16	55.24	**67.38**	65.02
BioASQ	78.93	72.21	**93.36**	93.14
BLURB score	80.09	77.11	**82.86**	81.88
Δ BASE model	-0.59	+0.15	+0.58	+0.37

Tableau 3 : COMPARATIF DE MODELES LARGE DE BERT SUR BLURB

Pour synthétiser : les modèles les plus efficaces sont donc les modèles les plus grands (*LARGE*), qui utilisent un vocabulaire biomédical dédié obtenu par un pré-entraînement sur des données biomédicales mais que l'ajout de données cliniques n'est pas forcément significatif.

[74] Tinn R., Cheng H., Gu Y.Fine-Tuning Large Neural Language Models for Biomedical Natural Language Processing. arxiv.org, 2021

5 MISE AU POINT SUR DEUX PROBLEMATIQUES TECHNIQUES QUI COMPLIQUENT LA TACHE

5.1 LE BIONER A LA RESCOUSSE DES FAIBLESSES DU CORPUS D'ENTRAINEMENT, MAIS UNE CHARGE DE TRAVAIL HUMAIN SUPPLEMENTAIRE

5.1.1 DEFINITION DU BIONER

Pour que l'application des techniques du TAL soient effectives, il faut qu'il préexiste un corpus d'entraînement riche et complet, par rapport à l'objectif qu'on s'est fixé. Créer ce corpus et chronophage et coûteux, car il nécessite une expertise fine et un *taguage* minutieux.

La spécificité des documents indexés dans IVAN, le fait que bon nombre d'entre-deux soient en français, le nombre de *labels* et la sélectivité des sujets font que passer par cette phase fastidieuse est dans notre cas d'étude, en l'état, nécessaire. Pour ce faire, il faut avoir recours à des sous-tâches de TAL que je vais exposer ci-dessous.

Couramment, dans le champs du TAL : la Reconnaissance d'Entité Nommée (REN ou NER en anglais pour *Named Entity Recognition*) est une méthode d'extraction d'unités d'information telles que des noms, comme ceux de personnes, d'organisations et de lieux, ainsi que des expressions quantitatives telles que l'heure, la date, les devises ou les pourcentages, à partir de données non structurées comme les textes.

En revanche, dans le domaine biomédical, les entités sont regroupées en classes telles que les gènes/protéines, les médicaments, les effets indésirables, les métabolites, les maladies, les tissus, les PSN (pour polymorphisme d'un seul nucléotide), les organes, les toxines, les aliments ou les chemins cliniques (*pathways*[75]). On parle alors de BioNER. Étant donné que l'identification des entités nommées est généralement suivie de leur classification dans une terminologie standardisée ou normalisée, elle est également appelée « reconnaissance et classification des entités nommées » (NERC en anglais pour Named Entity Recognition and Classification). Par conséquent, les deux termes, à savoir NER et NERC, sont fréquemment utilisés de manière interchangeable.

L'une des raisons pour lesquelles le BioNER est difficile est l'utilisation non standardisée d'abréviations, de synonymes, d'homonymes, plus généralement d'ambiguïtés langagières. On peut ajouter à ça le fait que bon nombre d'« entités » peuvent être constituées de phrases. Un exemple, est : « la condition neuropsychologique du syndrome d'Alice au pays des merveilles », qui, ici, nécessite la détection d'une chaîne de mots.[76]

[75] Dans les pays anglosaxons, décrit, pour une pathologie donnée, tous les éléments du processus de prise en charge en suivant le parcours du patient au sein de l'institution cf. https://www.has-sante.fr/upload/docs/application/pdf/2009-08/chemin_clinique_4.pdf Consulté le 16/11/2022

[76] Perera, N., Dehmer, M., Emmert-Streib F. Named Entity Recognition and Relation Detection for Biomedical Information Extraction, Front Cell Dev Biol, 2020

Dans le contexte biomédical, la reconnaissance d'entités nommées est souvent suivie de la détection de relation (RD en anglais pour *Relation Detection*, également connue sous le nom d'extraction de relation ou d'association d'entités). C'est-à-dire la connexion de diverses entités biomédicales les unes avec les autres pour trouver des interactions significatives qui peuvent être explorées plus avant.[77]
En analysant des millions d'articles de recherche biomédicale à l'aide d'approches informatiques, il est possible d'identifier des millions de ces associations pour créer des réseaux. Par exemple, l'identification des interactions des protéines permet la construction de réseaux d'interactions protéine-protéine. De même, on peut localiser des relations gène-maladie permettant de faire le pont entre l'information moléculaire et l'information phénotypique.[78]

5.1.2 L'EXTRACTION PROBLEMATIQUE DES ENTITES D'INTERET BIOMEDICAL

Plus précisément, la première étape implique le marquage des entités d'intérêt biomédical, comme le montre la figure, ci-dessous[79], avec l'exemple de phrase « Le gène BRCA1 provoque une prédisposition au cancer du sein et au cancer de l'ovaire ». Ici, les entités labélisées sont : « BRCA1 », « Cancer du sein » et « Cancer de l'ovaire ». Dans l'étape suivante, les relations entre ces entités sont déduites à l'aide de plusieurs techniques, telles que l'association indiquant des verbes comme illustré dans l'exemple. Ici, le verbe conjugué à la troisième personne « *causes* » (provoque) est identifié comme pointant vers une association possible. L'étape suivante, nous sert à distinguer la polarité de la phrase et le degré de la relation inférée. Par exemple, dans la phrase ci-dessus, la polarité est négative, c'est-à-dire qu'elle indique une relation défavorable entre le gène BRCA1 et la maladie labélisée et le degré (la vigueur) de la relation pourrait être extrait soit par le chemin le plus court dans l'arbre de dépendance de la phrase, soit par une simple distance de mot, comme indiqué dans l'exemple. Ce qui facilite l'exploration et la découverte des associations directes et des interactions indirectes.[80]

[77] Ibid.

[78]Ibid.

[79] Ibid.

[80] Ibid.

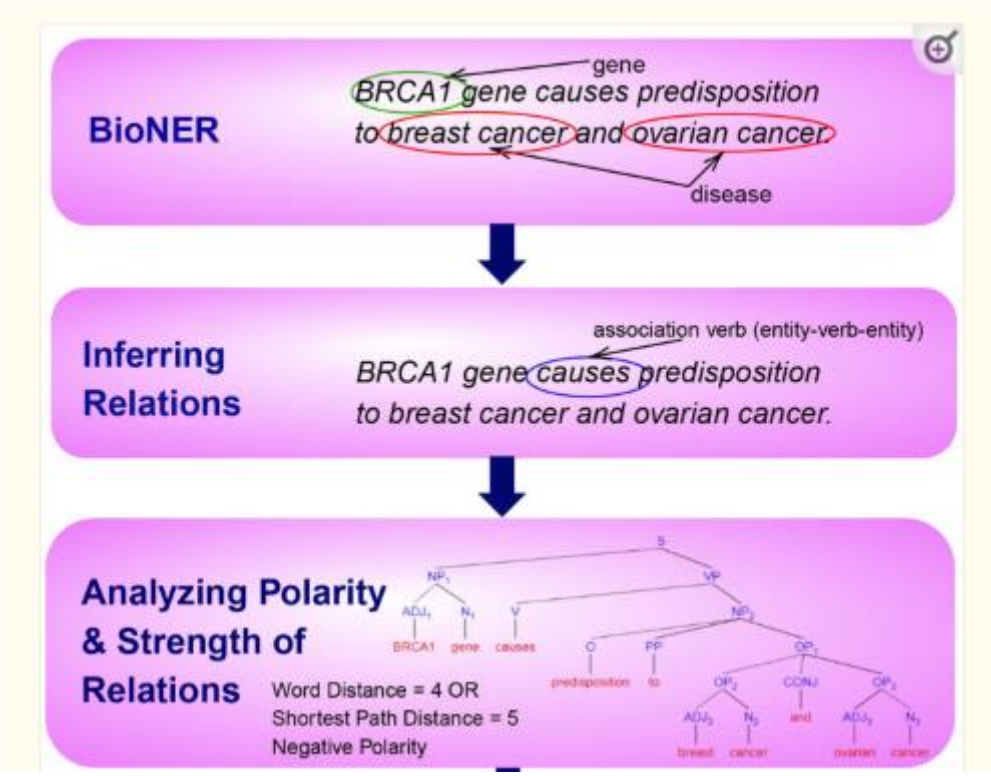


Figure 31: Les étapes d'une analyse de BioNER

Face au manque de *data* d'entraînement ou la généricité du vocabulaire du modèle, Il pourrait être judicieux (ici c'est une piste spéculative) de forcer le modèle à avoir un comportement vertueux, en limitant son degré de liberté dans la génération de ses représentations, par un marquage plus fastidieux certes, et là on fait face à un arbitrage entre autonomie d'apprentissage du modèle ce qui est la philosophie du *deep learning* et le f*eature engeneerin*g à la main (ce qui à l'opposé est plus proche de la philosophie de l'IA symbolique et des modèles à règles), mais permettant de combler le manque et d'avoir des résultats reproductibles de meilleur qualité.

Ces deux approchent dans la pratique sont complémentaires. J'en veux pour preuve qu'il existe de nombreux problèmes d'ambiguïtés sémantiques. Dans les passages suivant je vais en caractériser quelques un, qu'il me paraît judicieux d'avoir en tête, et qui à mon sens justifie un travail manuel d'annotation consciencieux pour pallier autant se faire que peut les angles morts du modèle.

Une première catégorie de confusions possibles pour le modèle, est l'utilisation des homonymes, de la métonymie, de la polysémie et des abréviations. En effet, alors que la plupart des termes dans le domaine biomédical ont une signification spécifique, il existe encore des termes, par exemple, pour les gènes et les protéines qui peuvent être utilisés de manière interchangeable. Comme GLP1R qui peut faire référence au gène ou à la protéine. Il faut une intervention humaine, donc un marquage pour résoudre de telles complications. Il y a aussi ces termes qui ont été utilisés pour décrire une maladie en termes profanes ou des médicaments qui

ont des noms de marque ambigus. Par exemple, des maladies comme « le syndrome d'Alice au pays des merveilles », le syndrome de Cotard qui en anglais se dit *Laughing Death* ou « le syndrome de l'accent étranger » et des noms de médicaments tels que Sonata, Yasmin, Lithium. Ce sont des candidats faciles à confondre pour des modèles n'ayant pas été pré-entraîné sur un corpus biomédical suffisant.
Et même dans ce dernier cas on peut faire face à un autre problème critique. Avec, par exemple le recours fréquent dans la littérature biomédicale d'abréviations aux significations ambiguës, telles que « CLD », qui pourrait faire référence à « *Cholesterol-lowering Drug* », « *Chronic Liver Disease* », « *Congenital Lung Disease* » ou « *Chronic Lung Disease* ».[81] Compte tenu des différences de signification et de classe, il est crucial d'identifier la bonne et d'apprendre au modèle à les contextualiser.

Alors que la plupart des problèmes ci-dessus sont le résultat du manque de nomenclature standard dans certains domaines biomédicaux, même les noms d'entités biologiques les plus standardisés peuvent contenir de longues chaînes de mots, de chiffres et de caractères de contrôle (par exemple *« 2,4,4,6- Tetramethylcyclohexa-2,5-dien-1-one », « spasme diaphragmatique épidémique transitoire »*)[82]. Ces entités nommées depuis longtemps rendent la tâche de BioNER complexe, ce qui pose des problèmes pour définir les limites des séquences de mots faisant référence à une entité biologique.

Ainsi, des définitions correctes des limites sont essentielles dans la phase d'entraînement du modèle, pour qu'il puisse capturer l'entité complète. L'une des solutions les plus couramment utilisées pour relever le défi de la capture de plusieurs mots consiste à utiliser un modèle de représentation *multi-segments* (SR) pour labéliser les mots d'un texte en tant que combinaison de « *Inside, Outside, Beginning, Ending, Single, Rear or Front* » en utilisant des normes comme le format IOB (pout Inside, Outside, Beginning). En voici un exemple issu du domaine général[83] :

“Alex **I-PER**
is **O**
going **O**
to **O**
Los **I-LOC**
Angeles **I-LOC**
in **O**
California **I-LOC”**

Dans cet exemple : « Alex » est tagué comme *Inside* et *Person*, « *is* », « *going* », « *to* » comme *Outside. « Los* » comme Inside, « *Angeles* » comme *Inside* et *Location, «* in » comme

[81] Ibid.

[82] Ibid.

[83] https://en.wikipedia.org/wiki/Inside%E2%80%93outside%E2%80%93beginning_(tagging) Consulté le 16/11/2022

Outside et « *California* » comme *Inside* et *location*. On voit qu'en plus de la labélisation, comme personne ou localité le modèle a appris à identifier pour chaque *token*, s'il « fait partie » d'une entité nommée et quel *token* est « en dehors » des entités nommées, ce qui lui permet de comprendre que « *Los* » et « *Angeles* » font partie de la même entité nommée de lieu.

Parfois la segmentation nécessaire dépend complétement du point de vue du documentaliste scientifique et de la finalité, du destinataire de la curation de l'article. Dans l'illustration, ci-dessous[84], on a un graphique qui illustre cette idée d'évaluation subjective de l'extraction correcte ou non de l'entité nommée. De gauche à droite l'échelle des valeurs possibles du plus strict au plus souple.

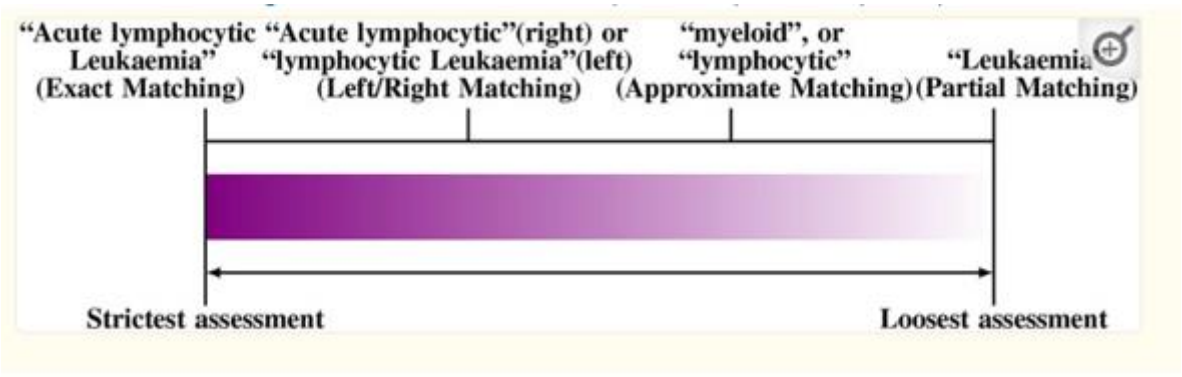


Figure 32 : Evaluation, plus ou moins stricte, de la validité de la reconnaissance d'une entité nommée

Pour finir cette série d'exemples : Les mêmes noms d'entités peuvent être écrits différemment dans différents articles. Par exemple : « *Lymphocytic Leukemia* » et « *Lymphoblast Leukemia* » sont synonymes, ce sont seulement des différences orthographiques britanniques et américaines).[85] Certains noms peuvent partager le même nom principal dans un article, comme dans « protéines 91 et 84 kDa » où ces entités imbriquées correspondent à « protéine 91 kDa » **et** « protéine 84 kDa », Dans ce cas la catégorisation doit tenir compte du contexte[86]. Pour résumer :

Pour réaliser une tâche de BioNER qualitative il faut développer un système complet pour capturer les entités nommées. Cela nécessite de définir les types des entités nommées, arbitrer des choix clairs pour déterminer des classes spécifiques pour les types des entités nommées et résoudre les problèmes sémantiques tels que la métonymie et les entités multi-classes. Enfin, il faut capturer les limites valides pour une entité nommée.

[84] Perera, N., Dehmer, M., Emmert-Streib F. Named Entity Recognition and Relation Detection for Biomedical Information Extraction, Front Cell Dev Biol, 2020

[85] Ibid.

[86] Ibid.

Une seconde étape, pourrait être l'extraction de relations entre des entités biologiques présentant un intérêt particulier pour la recherche médicale (par exemple, gène/maladie ou maladie/médicament).

Finalement, pour parfaire le système, il faut être en mesure de relever les défis syntaxiques et sémantiques propres au domaine biomédical et cruciaux pour qualifier le document qui nous intéresse.

Les principales étapes de BioNER comprennent : le prétraitement, le traitement des caractéristiques, la formulation/formation du modèle et le post-traitement cf. la figure ci-dessous[87].

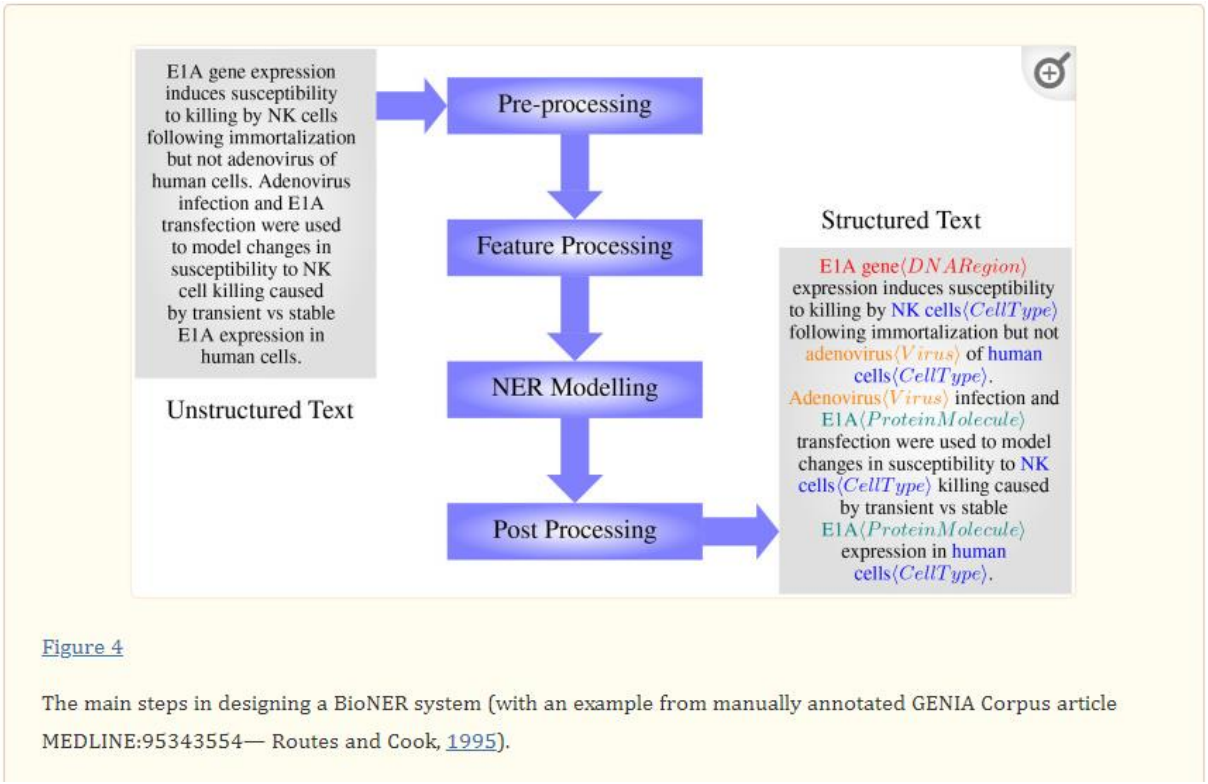


Figure 33 : Les étapes du processus de BioNER

Dans l'étape de prétraitement, les données sont nettoyées, *tokenisées* et, dans certains cas, normalisées pour réduire l'ambiguïté à l'étape de traitement des caractéristiques.

Le traitement des caractéristiques comprend différentes méthodes qui sont utilisées pour extraire les caractéristiques qui représenteront le plus les classes en question. Puis, il faut les convertir en une représentation appropriée pour appliquer la modélisation.

Ces étapes, grâce à des modèles basés sur BERT, sont largement simplifiées. Mais pour améliorer les représentations, les *Word embeddings*, comme nous l'avons montré, il est

[87] Ibid.

nécessaire d'entraîner le modèle à avoir une labélisation aussi précise que dans l'exemple ci-dessus et donc d'avoir recours à un travail patient mais je l'espère gratifiant d'annotation préalable.

5.1.3 CAS D'ANALYSE D'UNE NOTE CLINIQUE

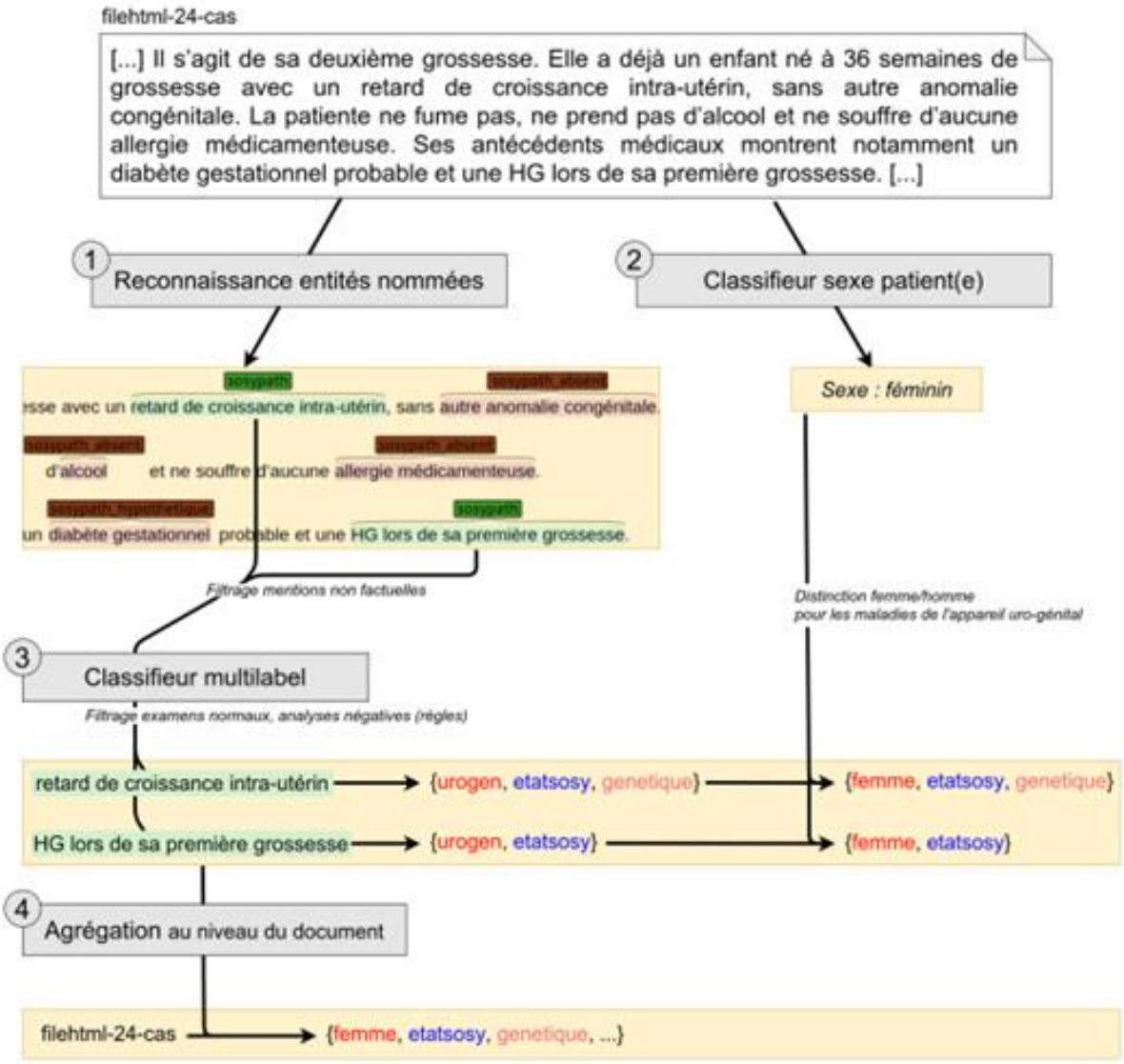


FIGURE 3 – Architecture générale du système

Figure 34 : Process d'analyse d'une note clinique par un système complexe de TAL

Un petit détour, par un cas d'analyse d'une note clinique[88] qui exploite de manière fine ce travail de forçage d'un modèle, ici CamenBERT (nous décrirons plus en détail ce modèle dans une partie dédiée), pour lui apprendre à tirer profit au maximum du langage naturel.

[88] Gérardin, C., Vaillant, P., Wajsbürt, P., *et al.* Classification multilabel de concepts médicaux pour l'identification du profil clinique du patient. *Traitement Automatique des Langues Naturelles,* 2021

Les entités nommées de cet exemple[89] consistent en deux catégories : la négation d'états pathologiques, en rouge, et au contraire, en vert, des états positivement symptomatiques. On voit donc dans cette exemple que la nature positive ou négative de ces affirmations médicales vient compliquer le travail de l'algorithme et qu'il faut filtrer les informations négatives et les écarter sous peine de fausser la classification finale. À cela s'ajoute, qu'il peut être nécessaire de déterminer le sexe quand un cas patient est évoqué dans le but d'évaluer, d'interpréter une situation médicale spécifique.

Il faut noter que comme dans le cas de la négation la détermination du genre se fait essentiellement à partir d'indices linguistiques.

L'exemple est intéressant car l'on voit que l'on part d'un paragraphe c0nséquent et non pas d'une simple phrase.

Concernant l'information médicale contenu : Est-il nécessaire de labelliser la classe de l'entité nommée (« symptôme », « Sosy » etc.) pour plus de robustesse ? Ou peut-on directement labelliser l'information avec un mot -clé issu du vocabulaire contrôlé, du thésaurus comme une « rubrique » pour rester dans le cadre de l'ICM ? Malheureusement on ne peut répondre à cette question que par l'expérimentation. Dans tous les cas on a ici, l'illustration d'une phase de REN préalable à une classification multi-label.

5.1.4 BIEN CHOISIR SON LOGICIEL D'ANNOTATION.

Ce travail préalable nécessaire est, hélas, chronophage, d'aucuns le jugeront prohibitif. Dans ces conditions, Il est impératif que l'outil ne soit pas un frein en lui-même, qu'il facilite au mieux la tâche de l'annotateur.

Dans cet article[90], les auteurs ont récemment fourni une revue approfondie des outils d'annotation, ils ont identifié une liste de critères qui, dans la pratique, rendent vraiment fonctionnels dans, leur usage, ces outils. Ces critères, si dans un premier temps on regarde leur aspect technique, doivent respecter ce cahier des charges :

- Ils doivent être accessible au plus grand nombre
- Hébergés sur le web
- Open source
- Mais aussi permettre une installation locale pour les documents sensibles comme les dossiers cliniques, ou les documents soumis au droit d'auteur.

[89] Ibid.

[90] Islamaj, R., Kwon, D. Kim, S., *et al*. TeamTat: a collaborative text annotation tool. Nucleic Acids Research, 2020

Concernant la nature des données traitées, ces outils doivent :

- Permettre de gérer les formats standards d'entrée et de sortie des documents et des annotations
- Permettre de traiter facilement les documents et données issus de PubMed.

Concernant les fonctionnalités, dans la pratique les utilisateurs veulent un outil qui gère :

- Les annotations multi-labels
- Les annotations relationnelles
- Les annotations au niveau global du document
- Les annotations avec une vue intégrale du document.
- Des annotations qui peuvent se faire collaborativement
-

En outre, un outil qui se positionnerait dans le haut du panier doit :

- Supporter plusieurs langues
- Prendre en charge les liens vers les ontologies et les thésaurus
- Permettre un contrôle qualité
- Permettre l'arbitrage entre les désaccords de plusieurs annotateurs.

Plus précisément en ce qui nous concerne, le point critique qu'il faut souligner : c'est l'intégration avec PubMed ou PubMed Central pour faciliter la récupération, l'analyse et le prétraitement des documents pour annoter un corpus du domaine biomédical

L'outil qui a été développé pour répondre exactement à ce cahier des charges et même à l'étoffer est
TeamTat[91], un outil d'annotation de texte : collaboratif, *open source*, basé sur le Web, étudié pour gérer la production de corpus annotés de haute qualité. En bref, TeamTat propose :

- Un support en texte intégral montrant le document dans son intégralité, y compris les graphiques, car ils font partie intégrante de l'annotation biomédicale manuelle comme nous l'avons énoncé précédemment (lors de l'analyse du travail des documentalistes scientifiques chez Novartis).
- Une intégration facile avec PubMed et PMC via BioC : un format simple pour partager des données textuelles et des annotations en vue d'une interopérabilité améliorée (créée par la communauté de recherche du *Text mining* biomédical)
- Une interface intuitive et conviviale permettant à tous les utilisateurs de revoir et d'analyser leurs annotations, de manière indépendante et collaborative.
- Un mécanisme d'évaluation et de gestion de la qualité avec la possibilité d'une supervision par un chef de projet.

Pour résumer TeamTat est un système tout-en-un avec un ensemble de fonctionnalités qui n'existaient pas dans les outils précédents.

5.2 LE PROBLEME DU BILINGUISME DES ARTICLES INDEXES DANS IVAN

[91] https://www.teamtat.org/ Consulté le 16/11/2022

« Au vu des récentes avancées en termes de performances et accessibilité, la traduction automatique neuronale apparaît comme la solution pour briser les barrières linguistiques freinant la circulation de l'information et des savoirs. Son rôle deviendrait donc déterminant dans un paysage à dominante anglophone comme la communication scientifique, qui pourrait ainsi se projeter vers un multilinguisme systématique. […] [Ces] perspectives, [de] la traduction automatique dans la communication scientifique, trouvent tout leur sens et montrent la nécessité de mettre en place des solutions *ad hoc* afin que ces technologies puissent réellement répondre aux besoins des communautés scientifiques. »[92] Comme le dit si joliment la linguiste, Fiorini Susana, dans cet article d'I2D.

Mais justement ça c'est le futur…Car, l'indexation automatique d'IVAN se heurte à un écueil de taille : les articles qui y sont indexés peuvent être aussi bien en anglais qu'en français, et la barrière de langue reste bien présente : le multilinguisme dans le domaine scientifique et, *a fortiori*, dans le biomédical reste une gageure.

En effet, tous les modèles dont nous avons parlé précédemment sont entraînés sur des corpus anglophones. Il existe une version de BERT multilingue qui couvre 104 langues dont le français[93]. Or, comme nous l'avons expliqué, la taille du vocabulaire de BERT se limite à 30 000 *tokens* et il doit le partager entre ces 104 langues ! Cela pose question quant à sa capacité à capturer des représentations décisives pour une tâche du domaine du TAL biomédical. Ce problème est renforcé par le fait que le modèle a été pré-entraîné exclusivement sur Wikipedia.

Il convient donc de faire appel à des modèles conçus par des spécialistes pour réviser ces outils et les adapter particulièrement aux domaines scientifiques français.

Nous allons présenter dans cette partie, des corpus français biomédicaux. Ces corpus utilisent un modèle conçu par l'INRIA dérivé de RoBERTa : la variante de BERT développée par les équipes de Facebook et considérée comme plus robuste.

CamemBERT a été entraîné sur un corpus générique mais optimisé pour le français et il est très efficient.

En effet dans un contexte où peu de corpus annotés français complets pour l'extraction d'entités médicales sont disponibles, une approche hybride peut être nécessaire. C'est à dire une approche combinant l'utilisation de connaissances spécialisées et l'adaptation de modèles de langues. Dans un article une équipe de chercheurs a mis l'accent sur l'effet du pré-entraînement d'un modèle de langue généraliste (CamemBERT) sur différents corpus.

Ils ont obtenu des résultats à partir du corpus QUAERO[94]. Ils ont montré que pré-entraîner un modèle avec un corpus spécialisé, même de taille réduite, permet d'observer une amélioration des résultats.

---

[92] Fiorini, S. L'intelligence artificielle au défi du multilinguisme : usages et perspectives de la traduction automatique neuronale dans la communication scientifique.*in* I2D

[93] https://huggingface.co/bert-base-multilingual-cased Consulté le 12/11/2022

[94] QUAERO est une compilation de deux corpus français annotés à dix types d'entités (Anatomie, Chimie et Médicaments, Dispositifs, Troubles, Zones Géographiques, Êtres Vivants, Objets, Phénomènes, Physiologie, Procédures). Ces deux corpus sont :

En combinant plusieurs approches ils ont pu gagner de 1 à 7 points de *F1 score* selon le corpus de test et la méthode.[95]

Dans un autre article, la solution présentée pour une tâche de classification *multi-label* d'un document (ici des notes cliniques issu du défi fouille de texte (DEFT) 2021[96] est aussi basé sur CamemBERT. Ils ont utilisé deux versions du modèle respectivement : le CamemBERT-large classique et le CamemBERT-large *fine-tuné* sur des articles biomédicaux français en accès libre, à partir de la base PMC Europe dont ils ont extraits 4000 articles. Ils ont noté que la faiblesse du volume de leur corpus ne permettait pas de conclure que dans ces conditions le *finetuning* pouvait être un facteur déterminant.[97] Cela est une limitation du *finetuning*, qu'il faut avoir à l'esprit, quand on le compare aux résultats obtenu sur un corpus anglophone à base d'articles issus de PubMed qui peuvent se chiffrer en million. Mais ce qui est intéressant, c'est qu'ils ont combiné plusieurs méthodes pour pallier ces limitations, dont une phase de reconnaissance d'entité nommées avec un spécificité : des labels doubles ; Pour un même label, il existe sa version « positive » et sa « négation » (par exemple un patient ne souffrant pas de tel symptôme). De plus, ils ont intégré un classifieur de genre. Ce système complexe est précieux car c'est l'illustration de la granularité jusqu'où il faut parfois aller pour extraire une information biomédicale de qualité.

Une autre *baseline* sur ce corpus français a été réalisée par une équipe qui a exploré des modèles de langage contextualisés pour la NER dans les textes biomédicaux français également dans le cadre du Défi Fouille de Textes. Leur meilleure approche a obtenu un *F1 score* de 66% pour les symptômes et les signes, et les catégories de pathologie. Pour les catégories anatomie, dose, examen, mode, moment, substance, traitement et valeur, ils ont obtenu un *F1 score* de 75%. Si l'on considère toutes les catégories, leur modèle a obtenu le meilleur résultat dans le cadre de ce défi, avec un F1 score de 72%. L'utilisation d'un ensemble de modèles de langages

---

– L'EMEA qui inclut de longs textes contenant des informations sur les médicaments commercialisés par l'Agence des médicaments.

– MEDLINE regroupe les titres des articles de recherche.

https://quaerofrenchmed.limsi.fr/ Consulté le 21/08/2022

[95] Le Clercq de Lannoy, T., Besançon, R., Ferret O., *et al.* Stratégies d'adaptation pour la reconnaissance d'entités médicales en français. 2022

[96] Le défi fouille de textes (DEFT) est une campagne d'évaluation annuelle francophone. Ainsi La première tâche du Défi fouille de textes 2021 a consisté à extraire automatiquement, à partir de cas cliniques, les phénotypes pathologiques des patients regroupés par tête de chapitre du MeSH-maladie. Il fallait identifier le profil clinique de patients décrits dans des cas cliniques.

[97] Gérardin, C., Vaillant, P. Wajsbürt, P., *et al.* Classification multilabel de concepts médicaux pour l'identification du profil clinique du patient. 2021

neuronaux s'est révélée très efficace, améliorant une base de référence du CRF[98] de 28% et un modèle de langage spécialisé unique de 4%[99].

Il existe un deuxième corpus très intéressant de référence, français, issu de l'entrepôt de données de l'Assistance Publique-Hôpitaux de Paris. Une équipe française a exploité un corpus de 21 millions de notes cliniques collectés d'août 2017 à juillet 2021[100]. L'algorithme utilisé a effectué des tâches de reconnaissance d'entités nommées et ont obtenu un *F1 score* de 91%. L'algorithme en question est CamemBERT. Ils montrent quant à eux que le *finetuning* de modèles généralistes, tels que CamenBERT, sur des corpus spécialisés français améliore leurs performances pour les tâches de TAL biomédicales. Leurs résultats suggèrent en revanche, que l'entraînement à partir de zéro n'induit pas de gain de performance statistiquement significatif par rapport au réglage fin (*finetuning*) mais là aussi cela peut être biaisé par la faiblesse du corpus de pré-entrainement, corpus qu'ils n'ont pas, par ailleurs, libéré.

Actuellement, deux modèles linguistiques sont spécialisés en langue française : FlauBERT et CamemBERT.

Une étude a été conçue pour déterminer quelle combinaison de modèle de langage et d'architecture de réseau de neurones était la meilleure pour des tâches de prédiction par un chatbot à partir d'un corpus français de cas cliniques. Les comparaisons ont montré que FlauBERT était plus performant que CamemBERT quelle que soit l'architecture de réseaux utilisée et que les architectures complexes n'amélioraient pas significativement les performances par rapport aux architectures simples quel que soit le modèle de langage.

Ainsi, dans le domaine médical, les résultats appuient la recommandation de FlauBERT avec une architecture de réseau linéaire simple.[101]

---

[98] CRF (Conditional Random Field) est un modèle discriminant probabiliste qui un large éventail d'application en Machine Learning. En termes simples, le modèle discriminatif modélise la frontière de décision entre différentes classes. L'exemple courant d'un modèle discriminatif est la régression logistique qui maximise les estimations de vraisemblance. Très tôt utilisé dans le Machine Learning, le CRF est donc un candidat idéal pour mesurer les progrès des nouvelles approches dans un test comparatif. Pour plus de détail sur ce modèle on peut se référer à la page du Wikipédia français : https://fr.wikipedia.org/wiki/Champ_al%C3%A9atoire_conditionnel#:~:text=Les%20champs%20al%C3%A9atoires%20conditionnels%20(conditional,plus%20g%C3%A9n%C3%A9ralement%20en%20apprentissage%20statistique Consulté le 23/08/2022

[99] Copara, j., Knafou,J., Naderi, N., *et al.* Contextualized French Language Models for Biomedical Named Entity Recognition. 2022

[100] Dura, B., Jean, C., Tannier, X., *et al.* Learning structures of the French clinical language: development and validation of word embedding models using 21 million clinical reports from electronic health records. arxiv.org, 2022

[101] Blanc, C., Bailly, A. Francis, E. FlauBERT vs. CamemBERT: Understanding patient's answers by a French medical chatbot. Artificial Intelligence in Medicine, Volume 127, 2022

6 SOLUTIONS POSSIBLES

6.1 SOLUTION SUR MESURE ET FAITE MAISON

La meilleure solution pour une indexation automatique des articles d'IVAN est un assemblage en un même système complexe de plusieurs composants que je vais exposer ci-dessous.

Tout d'abord, voici le schéma[102] du système que j'ai imaginé pour solutionner la problématique de classification *multi-label* au cœur de notre sujet de mémoire. Mais ce n'est qu'une illustration pour concrétiser en schématisant une solution possible d'architecture.

[102] Schéma réalisé par mes soins.

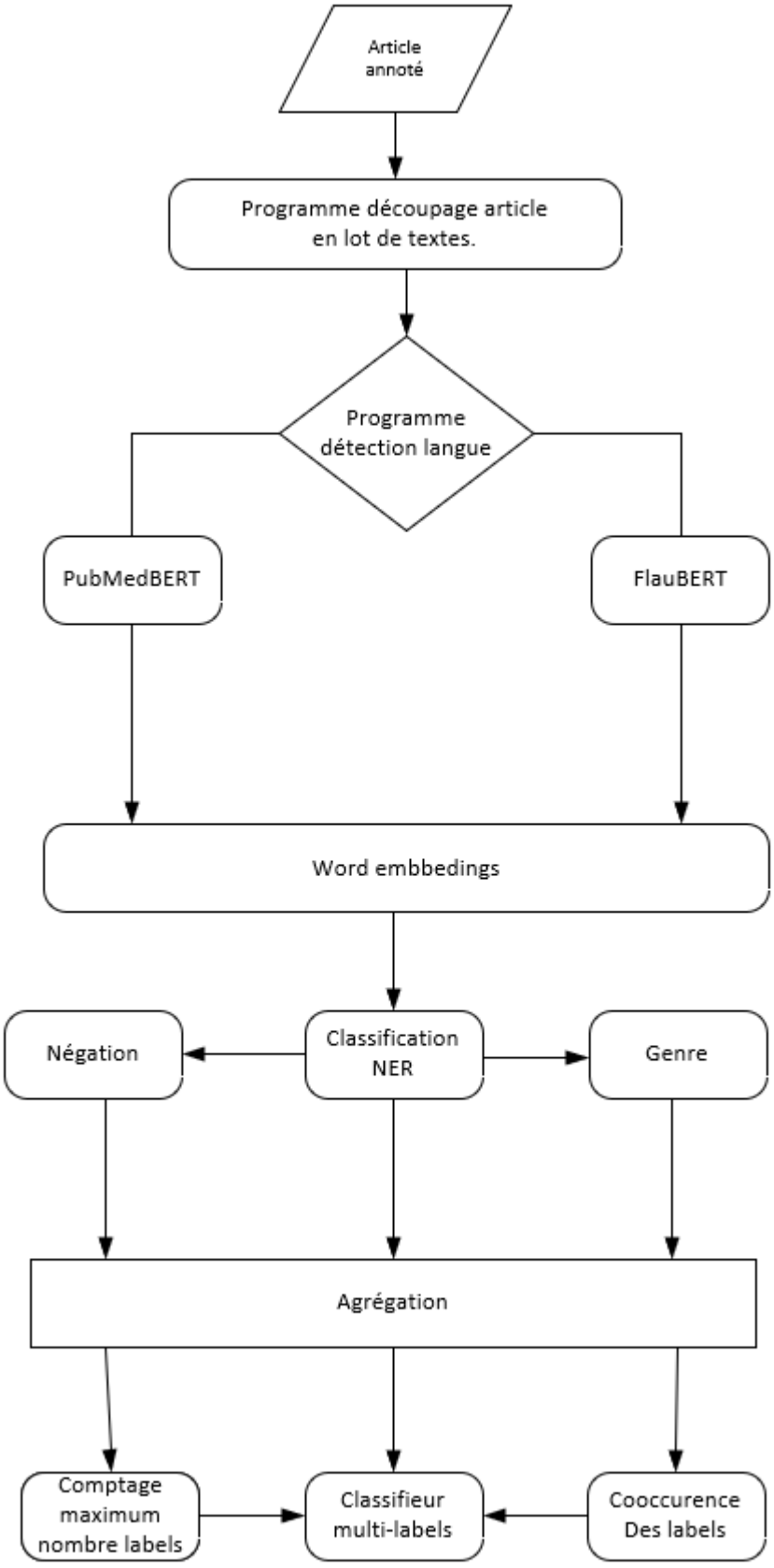


Figure 35 : Architecture du système DocICM

La première brique est un programme qui utilise la librairie python *langdetect* qui permet de détecter la langue de l'article et de choisir le modèle adéquate pour traiter l'article. Puis une phase de prétraitement qui divise l'article en plusieurs lots. Le premier lot est constitué de la concaténation du titre et de l'abstract, puis les lots qui suivent sont des parties arbitraires tronquées à 512 tokens, séquentiellement jusqu'à épuisement du texte. Ces lots de textes vont nourrir le modèle choisi :

- PubMedBERT pour les articles en anglais
- FlaubERT pour les articles en français

Les prédictions de chaque lot de texte vont être agrégés pour produire un résultat final de labélisation pour chaque document.[103]

Cette seconde brique, dans les deux cas, peut être décomposé ainsi :

Les classifieur identifiés comme BERT sur le schéma correspondent, bien sûr, soit à classifieur appartenant au modèle PubMedBERT si la langue anglaise est détectée soit à FlauBERT si le français a été détecté.

Les classifieurs *Négation* et *Genre*. Sont là, à titre d'exemple, pour illustrer comment on peut raffiner le modèle en injectant de l'information linguistique non biomédical. Ils peuvent être ignorés, si on les juge inutiles. Ils peuvent, également, être remplacés par d'autres items qui auraient été jugés plus pertinent.

Par exemple déterminer la négation, peut aider à caractériser un effet indésirable et le genre le profil clinique ou la population concernée, ciblée.

Il en va des même pour les items « Comptage maximum nombre label » et Cooccurrence » » des labels qui ont été empruntés aux papiers étudiés plus en amont dans l'étude.[104]

[103] Pour plus d'explication pour cette technique se référer à l'article de blog qui me l'a inspiré : https://medium.com/@armandj.olivares/using-bert-for-classifying-documents-with-long-texts-5c3e7b04573d, Consulté le 16/11/2022

[104] Ils ont été empruntés, respectivement au modèle ML-NET et au modèle LitMC BERT

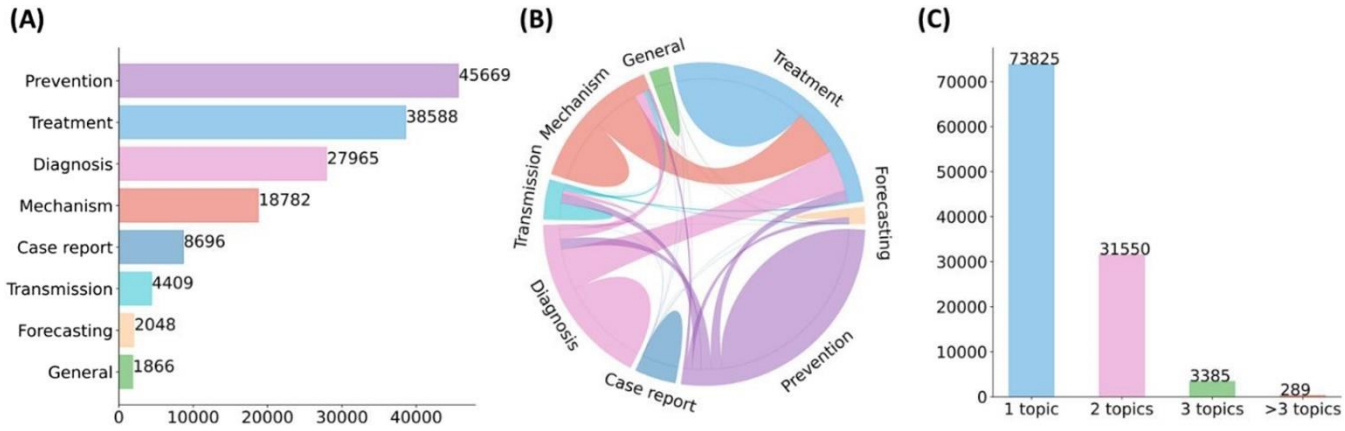


Figure 36 : La distribution des topics annotés dans LitCovid

La pertinence de permettre la détection de la cooccurrence des *labels* et de contrôler le nombre de ces mêmes *labels* est illustré par les graphiques ci-dessous[105].

Premier point, comme on peut le voir sur le graphique B, les *labels* sont plus ou moins corrélés et peuvent donc émerger ensemble. Ce fait est généralisable à toute la littérature biomédicale mais n'est pas systématique. Il peut donc paraître intéressant, d'apprendre au modèle à repérer ces cooccurrences pour obtenir des prédictions plus cohérentes et une qualification du document de meilleure tenue.

Deuxième point, il y'a des chances pour que la distribution des *labels* de l'ICM soit hétérogène, déséquilibrée. Même si le cas de LitCovid n'est qu'un aperçu, c'est une propriété courante (dans les cas de classification *multi-label* de document biomédicaux) que j'ai rencontrée. Il peut donc être judicieux, pour plus de robustesse, d'intégrer un mécanisme de régulation du nombre des *labels* de manière intelligente pour cadrer le modèle et qu'il sache où donner de la tête.

Last but not least, j'ai imaginé une solution complète et intégrée, au-delà de la seule indexation automatique, qui reposerait sur six couches superposées[106] :

[105] Chen, Q., Du, J., Allot, A., Lu, Z. LitMC-BERT: transformer-based multi-label classification of biomedical literature with an application on COVID-19 literature curation. IEEE/ACM Trans Comput Biol Bioinform, 2022

[106] Schéma réalisé par mes soins

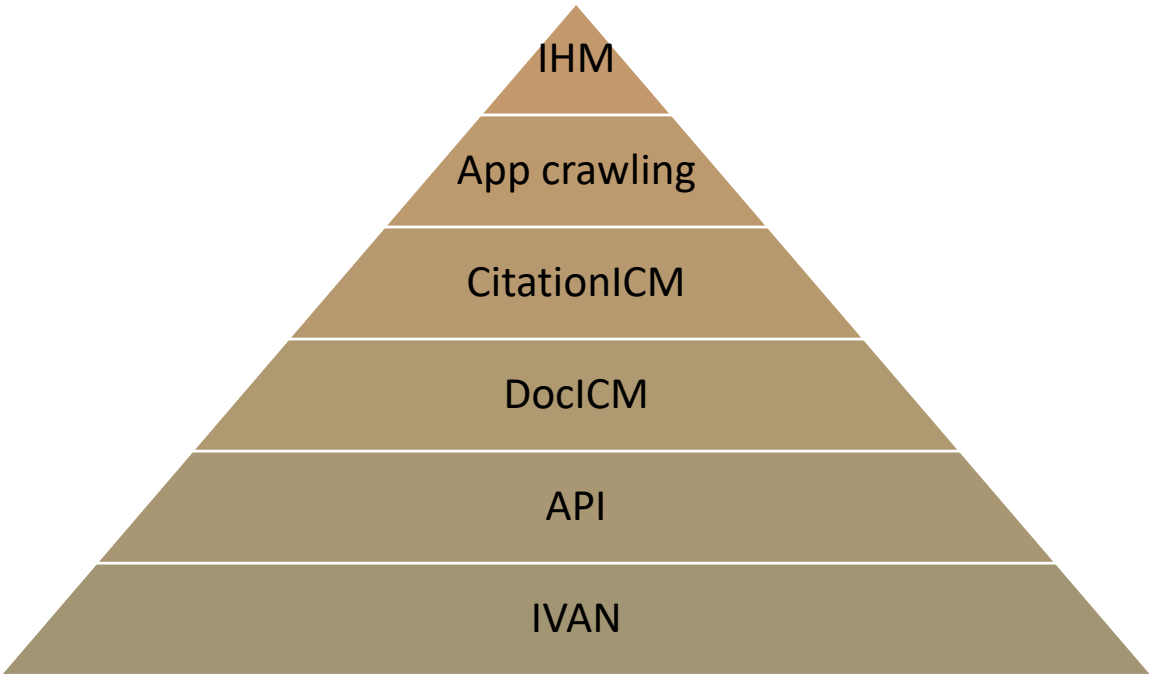


Figure 37 : Schéma d'un solution intégrée pour l'indexation automatique dans IVAN

1. la première couche : l'IHM, c'est à dire l'Interface Homme Machine. Une interface pensée pour être la plus ergonomique possible et fonctionnelle, selon les dernières avancées de l'UX, l'*User Experience*, avec des allez retours et des tests continus entre les documentalistes et le développeur *front end*, pour un usage optimal.

2. L'*App crawling* serait une application, faite maison, de veille, en amont, qui permettrait une présélection des articles, pertinents, parmi la masse du catalogue de l'ICM, tout en respectant les contraintes juridiques et techniques liées à chaque éditeur.

3. CitationICM : Un modèle entraîné sur le même corpus « maison » pour extraire les métadonnées bibliographiques qui seraient en sortie formattées selon le style de l'ICM et avec une marge d'erreur extrêmement faible (ce qui n'est pas le cas des outils utilisé jusqu'à présent).

4. DocICM : La brique fondamentale, le modèle permettant l'extraction des métadonnées par rapport au thésaurus de l'ICM, que nous avons schématisé en début de cette partie.

5. L'API, pour Interface de Programmation d'application. L'API est « une solution informatique qui permet à des applications de communiquer entre elles et de s'échanger mutuellement des services ou des données »[107]. C'est elle qui permettra l'indexation automatique dans IVAN.

6. Enfin, IVAN, la base de connaissance, qu'on ne présente plus !

6.2 DEUX EXEMPLES DE PRESTATAIRES EXTERNES A NOVARTIS QUI POURRAIENT APPORTER UNE SOLUTION.

6.2.1 KAIRNTECH[108]

KAIRNTECH c'est une société qui intervient dans le domaine du TAL en utilisant spécifiquement, les dernières technologies à l'état de l'art dans les réseaux de neurones : l'apprentissage actif (*Active Learning*) et l'apprentissage par transfert (*Transfer Learning*). Elle permet de gérer et d'analyser de grandes quantités de documents, de mettre à jour des vocabulaires métiers et des bases de connaissance.

La plateforme est accessible aux non-développeurs, c'est une plateforme d'IA « No-Code ». C'est donc une aide précieuse accessible à des non informaticiens. En effet, elle a une approche qui se veut simple et rapide du début à la fin du projet. Aucune compétence en programmation n'est nécessaire, chaque étape est pensée pour mettre au cœur de la plateforme l'expertise du professionnel de l'information. C'est lui qui à la main sur l'outil pour valoriser et exploiter aux mieux son domaine d'expertise. Cette approche *Plug and Play*, veut économiser de longues heures d'un travail fastidieux et chronophage en prenant en charge la complexité technique. Tout en garantissant l'automatisation du traitement documentaire pour la tâche qu'on s'est fixé de mettre en œuvre.

Les étapes d'un projet avec Kairntech prévoient :

1. L'importation du corpus de documents et du vocabulaire d'amorçage.

2. Un annotateur de texte pour annoter le corpus de documents avec les termes du vocabulaire d'amorçage.

[107] https://www.journaldunet.com/management/efficacite-personnelle/1515287-connaissez-vous-ces-sigles-informatiques/1515299-api, consulté le 13/11/2022

[108] NIBART, V. Découvrir et enrichir des connaissances à partir de l'analyse de documents grâce à l'intelligence artificielle : l'exemple de la plateforme KAIRNTECH. *in* I2D, op. cit.

3. Un contrôle qualité du corpus annoté, pour l'amender manuellement ou de manière assistée. Le corpus annoté devient alors un corpus d'apprentissage.

4. L'apprentissage automatique avec des moteurs de *Machine Learning* et de *Deep Learning* classiques, mais la plateforme permet, aussi, la génération de *Word embeddings* avec des modèles de type BERT. Ces moteurs sont fournis par la plateforme et sont mis à jour régulièrement. Un point intéressant est que l'utilisateur peut expérimenter et évaluer ces différents algorithmes sur son corpus cible.

Maintenant la question clé : Est-ce que cet outil n'est pas trop *mainstream ?* Est-ce qu'il est adapté et assez souple pour répondre à notre problématique et à toute ses contraintes ? Et si oui pour quel coût supplémentaire ?

6.2.2 QWAM[109]

QWAM *Content Intelligence* est un des acteurs français principaux du TAL appliqué aux entreprises qui a l'habitude de travailler avec des grands comptes. L'entreprise et ses outils excellent dans la génération de métadonnées, l'extraction d'informations et de relations, le classement et l'organisation de l'information.

Ils sont spécialisés sur les données de type fichier et documents textes comme les contrats, les rapports, les articles des médias généraux. Ils produisent des applications métiers et documentaires.

Ils sont en mesure de concevoir tout une gamme d'applications métier.

Mais, ils ont aussi un produit phare : Text Analytics, qui grâce à la sémantique, au *Big Data*, à l'intelligence artificielle, le *Deep Learning* et le *Machine learning*, avec une spécialisation historique sur le NLP (TAL), leur a permis de développer une expertise sur l'extraction de métadonnées et l'analyse de grande masse de textes.

Par exemple, l'outil est capable d'extraire, je les cite :

- Les entités nommées générales : Sociétés, personnes, lieux etc.
- D'autres entités comme : Les produits, les procédés, les composants, les substances, les molécules, Les maladies.

[109] QWAM Text Analytics, *l'Intelligence Artificielle (IA) et la sémantique pour mieux valoriser les contenus éditoriaux,* La Presse au Future, Février 2021, Consulté le 14/11/2022.

- Les concepts généraux thématiques (sur 15 thèmes)
- Les concepts spécialisés (propriétés mécaniques ou chimiques…, les opérations…
- Les relations standards,
- Des relations sur mesure.
- Enfin, les sentiments.

Leurs interfaces se déclinent en tableaux de bord, allant du nuage de mots aux facettes ou filtres pour moteurs de recherche.

Les usages, pour ce qui nous concerne, sont les usages autour des informations métiers, des moteurs de recherche et bases de connaissance, avec une analyse de la littérature scientifique et de la documentation technique qui permet une extraction d'indicateurs et de relations techniques ou scientifiques.

L'accès aux applications métiers peut se faire en mode *Saas*, c'est-à-dire sur une plateforme d'hébergement QWAM située en France ou par installation, *On Premise*, c'est-à-dire chez le client avec une intégration au SI de l'entreprise.

Une de leur force est l'accompagnement par des experts tout au long du projet, puis un suivi et une maintenance avec même la possibilité de faire évoluer l'outil au fil de l'eau.

C'est cet aspect qui me parait très intéressant dans le cadre de l'automatisation automatique de l'indexation dans la base de connaissance IVAN. Car ce qui ressort de l'étude que nous avons menée est le besoin de construire un outil sur mesure, du fait des contraintes que nous avons dégagé et la spécificité du domaine biomédical qui n'est pas leur cœur de métier, mais, qui au vu de leur expertise technique et de leur expérience auprès de grands groupes, devrait être dans leurs cordes.

Bien sûr un tel projet d'envergure, avec un accompagnement personnalisé, doit avoir un coût conséquent, que je ne suis malheureusement pas en capacités d'évaluer.

CONCLUSION

L'objet de ce mémoire était de réaliser l'étude de faisabilité de l'indexation automatique ou semi-automatique de la base de connaissance IVAN.

La base de connaissance IVAN est une base centralisant des articles scientifiques du domaine biomédical sélectionnés par les documentalistes scientifique de l'ICM : service de

documentation hébergé au siège de Novartis France. Ces articles sont enrichis de métadonnées à partir d'un thésaurus, ou plutôt un vocabulaire contrôlé biomédical, qui leur est propre. Une fois enregistrés, dans la base, ces articles sont diffusés aux collaborateurs de l'ICM à l'aide d'alertes que ces derniers ont paramétrées selon leurs besoins ou à partir d'une *newsletter* « l'Info Comme On l'M ».

Ce travail de curation est chronophage. En plus, du travail de veille fait en amont, qui est en lui-même conséquent, le processus est alourdi par la double saisi des métadonnées. Elles sont une première fois, pour partie, saisies dans une grille Excel conçue pour cet effet, puis ressaisies par une de leurs collègues dont c'est la principale tâche.

Une première prospective a eu lieu, à l'initiative de ma tutrice de stage, pour tenter d'alléger ce process d'indexation. Hélas, après étude l'outil retenu : Curebot (un logiciel de veille automatisée collaboratif) s'est révélé peu adapter aux besoins réels de l'ICM. Deux écueils majeurs ont été mis au jour :

- Le premier est la contrainte juridique : bon nombre des éditeurs auxquels l'ICM est abonné n'autorise pas le *crawling* sur leur site. Or cette pratique, qui vise à scanner le site pour repérer d'éventuelles mises à jour, est la base programmatique de ce type de logiciels.

- La deuxième contrainte est technique. En effet, ces logiciels qui scrutent des sites afin de sélectionner les derniers articles pour venir alimenter des dossiers paramétrés à l'aide de requêtes de recherche prédéfinies, le font en exploitant les flux RSS de ces sites. Or, ces derniers, pour la majorité du catalogue de l'ICM, n'ont pas implémenté cette technologie.

D'autre part, la collègue en charge de l'indexation finale dans IVAN utilise un outil pour extraire les métadonnées bibliographiques de ces articles. Or après usage, il s'est avéré que cet outil générait systématiquement des erreurs dans le formatage de ces données.

Donc, à ce stade, il est clair qu'on ne peut parler d'indexation automatique et qu'il serait plus adéquat d'envisager une indexation semi-automatique, cette partie en amont de la veille et du pré-remplissage des métadonnées bibliographiques, nécessitant, hélas, d'être fait à la main.

Il a donc été envisagé de « débroussailler le chemin » vers l'automatisation de l'indexation des métadonnées biomédicales de ces articles à l'aval : c'est-à-dire après le travail de veille par les documentalistes scientifiques. C'est là que j'interviens.

Premier point positif, le tout nouveau cadre réglementaire européen et français nous autorise à utiliser des solutions d'intelligence artificielle à des fins de fouille de textes à la seule condition que les droits de ces articles aient été au préalablement, légalement, acquis et que ces articles aient été supprimés des serveurs après l'opération de *datamining*.

Le second point positif est que cette problématique bénéficie du contexte bouillonnant de l'indexation des termes MeSH au profit de MEDLINE. La traction de la recherche en ce

domaine à pour première conséquence que la classification *multilabel* dans le domaine biomédical est bien documentée et porte sur des outils d'intelligence artificielle à l'état de l'art.

Troisième point positif est que depuis peu l'intelligence artificielle a connu l'essor du TAL ou NLP. C'est-à-dire que l'application des dernières avancées du *Deep Learning* : les *Words embeddings*, l'architecture *transformer* et les modèles pré-entraînés comme BERT, ont révolutionné le champs du TAL.

Ces modèles BERT ont été acclimatés au domaine biomédical. Et dans notre étude nous avons pu souligner les excellentes performances du modèle PubMedBERT : pré-entraîné exclusivement sur des données issus du domaine biomédical, en l'occurrence les abstracts PubMed.

Malheureusement, là on bute contre le premier problème : IVAN n'est pas implémenté pour intégrer les abstracts. Or, la plupart de ces modèles biomédicaux ont été entraîné à travailler à partir d'*abstracts*. Car, le modèle BERT, originel, a une limitation inhérente : c'est d'accepter une séquence de 512 *tokens* maximum à la fois.

En conséquence, une de mes premières recommandations serait de changer cette implémentation pour les futurs articles.

Un second problème est que l'information biomédicale est particulièrement complexe à saisir pour une machine. Outre, toutes les ambigüités langagières, pour avoir de bonnes représentations (*Word embeddings*), le modèle doit être en mesure de segmenter le texte de manière pertinente. Il peut donc être nécessaire d'avoir recours à des sous tâches du TAL biomédical : le BioNER et l'extraction de relations. Or ces tâches nécessitent un marquage minutieux des articles à fin d'en baliser l'information et de l'extraire. Ce *taguage* peut se faire avec l'outil adéquat et je recommande à cet effet le logiciel TeamTat.

Un troisième problème est le bilinguisme des articles indexés dans IVAN. En effet, un grand nombre d'entre eux sont en français. Or, la plupart de la recherche sur ces modèles biomédicaux est anglosaxonne. De plus les *datasets* d'articles biomédicaux se comptent en millions en littérature anglaise et en milliers pour le français. Ce pourquoi, il n'existe pas d'équivalent francophone au modèle PubMedBERT. Mais, heureusement, il existe des modèles français, génériques, robustes comme CamemBERT et FlauBERT. Je recommande FlauBERT car il apparaît comme le plus prometteur.

Si je tire un bilan des pour et des contre : je suis, à ce stade, obligé de répondre par la négative à la question de la faisabilité de l'indexation semi-automatique dans IVAN. Bien sûr, techniquement, c'est possible. Mais je ne le recommande pas. En effet, à mon avis c'est prématuré. Cela entrainerait un surcharge de travail de la part des documentalistes scientifiques que je juge prohibitive. Or, le but était de les soulager ! Pas de les accabler de travail !

En revanche, je pense que le champs que nous avons étudié évolue à la vitesse lumière ! Et nous avons dressé les premiers jalons pour une résolution future. En effet, il y'a des points

qui sont de bonne augure, dans les tendances actuelles : les *transformers* se généralisent pour devenir les fondations, non seulement dans le TAL, mais également dans tous les champs du *Deep Learning.* C'était important de les étudier, maintenant, nous avons de bonnes bases. Si la question se repose dans le futur, on saura d'où partir. Deuxième tendance, actuelle, qui peut avoir un impact déterminant sur cette question de faisabilité. Les grands groupes comme Google et Meta font une priorité de la traduction de plus de langues possibles. Avec un objectif de 1 000 langues traduites par leurs modèles, d'ici quelques années. Ces efforts de recherche peuvent ruisseler et résoudre la question de la langue telle qu'elle se pose pour IVAN.

Les solutions « maison » que nous proposons, que ce soit DocICM ou bien la solution intégrée, telle que nous l'avons imaginée dans sa globalité, au-delà de l'exercice, peuvent aider, et je le pense humblement, à se poser les bonnes questions quant aux problèmes de conceptions techniques et humains, plutôt qu'à établir une solution qui est encore à venir.

Enfin les prestataires identifier peuvent être des alliés solides dans la mise en place d'une telle solution.

Last but not least, il est intéressant de remarquer que dans le cadre de la restructuration actuelle «*Transforming for Growth*» : « l'info méd » et l' ICM pourraient apporter une contribution non négligeable à son développement en dynamisant son offre et en enrichissant sa R&D.

« L'info méd » est actuellement sous exploitée et pourrait apporter beaucoup plus de valeur ajoutée dans toute la chaine de valeur de Novartis du local au global.

Il s'agit de promouvoir une intelligence biomédicale qui associe l'humain et la technologie de façon vertueuse et responsable.

Les métiers de la documentation, loin de l'image parfois traditionnelle qu'on leur prête, peuvent être aux avant-postes de nombreuses innovations technologiques.

BIBLIOGRAPHIE

Notice : Bibliographie comprenant les principales ressources qui m'ont aidé à documenter mon mémoire. Ces références bibliographiques apparaissent dans l'ordre chronologique : une fois, à partir de la première occurrence dans le corps du mémoire. Elles sont classées par grande partie du mémoire. J'ai utilisé la norme ISO 690

INTRODUCTION

- Archimag.com. Intelligence Artificielle : Une veille augmentée ? N° 343, Paris, 2021. Groupe SERDA, 2008- , Mensuel. ISSN 2260-166X. Également disponible en ligne à l'adresse : < https://www.archimag.com/veille-documentation/2021/05/07/intelligence-artificielle-veille-augmentee > [Consulté le 22/08/2022]
- Chastenet de Géry G. Le knowledge management : Un levier de transformation à intégrer. De Boeck Supérieur, 2018
- The Economist. AI'S NEW FRONTIER.11-17, juin 2022
- Zhao S., Su C., Lu Z., et al. Recent advances in biomedical literature mining. Brief Bioinform, 2021

PREMIERE PARTIE : CADRE DE TRAVAIL ET ETAT DES LIEUX

- Novartis. Annual Report, 2021. Disponible à l'adresse : <https://www.novartis.com/sites/novartis_com/files/novartis-annual-report-2021.pdf > [Consulté le 20/11/2020]

DEUXIÈME PARTIE : ELÉMENTS DE CONTEXTE

- Commission européenne. COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS. A European strategy for data. Bruxelles, 2020. Disponible à l'adresse : < https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066 >[Consulté le 21/08/2022]
- DIRECTIVE (UE) 2019/790 DU PARLEMENT EUROPÉEN ET DU CONSEIL du 17 avril 2019 sur le droit d'auteur et les droits voisins dans le marché unique numérique et modifiant les directives 96/9/CE et 2001/29/CE. Disponible à l'adresse :< https://eur-lex.europa.eu/legal-

content/FR/TXT/PDF/?uri=CELEX:32019L0790&from=FR > [Consulté le 17/11/2022]

- I2D - Information, données & documents 2022/1 (n° 1) : L'INTELLIGENCE ARTIFICIELLE. A.D.B.S. Trimestriel.114p. ISSN en ligne : 2431-3467
- INSERM. Le MeSH bilingue. Information Scientifique et Technique. Disponible à l'adresse :< https://mesh.inserm.fr/FrenchMesh/ > [Consulté le 21/08/2022]
- Kublik V. EU/US Copyright Law and Implications on ML Training Data. Valohai, 2019 [En ligne].Disponible à l'adresse : <https://valohai.com/blog/copyright-laws-and-machine-learning/ > [Consulté le 21/08/2022]
- Medline 2022. Initiative : transition to automated indexing. NLM Tech Bull, 2021, (443):e5. Disponible à l'adresse : <https://www.nlm.nih.gov/pubs/techbull/nd21/nd21_medline_2022.html> [Consulté le 21/082022]
- ORDONNANCE n° 2021-1518 du 24 novembre 2021 complétant la transposition de la directive 2019/790 du Parlement européen et du Conseil du 17 avril 2019 sur le droit d'auteur et les droits voisins dans le marché unique numérique et modifiant les directives 96/9/CE et 2001/29/CE. Disponible à l'adresse :<https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000044362034> [Consulté le 14/11/2022]

TROISIÈME PARTIE : QUALIFICATION DE LA PROBLÉMATIQUE TECHNIQUE

- Chougule, A. Thoughts on NLP in biomedical domain. Linkdin, 2021. Disponible à l'adresse :< https://www.linkedin.com/pulse/thoughts-nlp-biomedical-domain-akshay-chougule > [Consulté le 21/09/2022]
- Goodfellow, I., Bengio, Y., Courville A. Deep Learning. MIT Press, 2016. 800 p. ISBN : 9780262035613
- Howard, J., Gugger, S. Deep Learning for Coders with Fastai and PyTorch. O'Reilly Media, 2020. 350 p. ISBN : 97814920455261 Ibid.

QUATRIEME PARTIE : BENCHMARK DES METHODES ALTERNATIVES

- Chen, Q., Du, J., Allot, A., Lu, Z. LitMC-BERT: transformer-based multi-label classification of biomedical literature with an application on COVID-19 literature curation. IEEE/ACM Trans Comput Biol Bioinform, 2022. Disponible à l'adresse : < https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9647722/> [Consulté le 20/11/2022]
- Du, J., Chen, Q., Peng, Y., Xiang, Y., et al. ML-Net : multi-label classification of biomedical texts with deep neural networks. Journal of the American Medical Informatics Association, Vol. 26, n° 11, 2019. p. 1279–1285
- Kalyan, K.S., Rajasekharan, A., Sangeetha S. AMMU : A Survey of Transformer-based Biomedical Pretrained Language Models. Journal of Biomedical Informatics,

Vol. 126, 2022. Disponible à l'adresse : < https://arxiv.org/pdf/2105.00827.pdf > [Consulté le 20/11/2022]

- Gu, Y., Tinn, R., Cheng, H., Lucas,M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon H. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Trans. Comput. Healthcare, 2021, n°1. 23 p. Disponible à l'adresse <https://arxiv.org/pdf/2007.15779.pdf> [Consulté le 20/11/2022]
- Lewis, P.S.H., Ott, M., Du, J., et al. Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. openReview.net,2020. Disponible à l'adresse : < https://aclanthology.org/2020.clinicalnlp-1.17.pdf> [Consulté le 20/11/2022]
- Rae, A.R., Mork, J. G., Demner-Fushman, D. A Neural Text Ranking Approach for Automatic MeSH Indexing. National Library of Medicine, 2021. Disponible à l'adresse : < https://ceur-ws.org/Vol-2936/paper-22.pdf > [Consulté le 20/11/2022]
- Tinn R., Cheng H., Gu Y. Fine-Tuning Large Neural Language Models for Biomedical Natural Language Processing. arxiv.org, 2021 Disponible à l'adresse : < https://arxiv.org/pdf/2112.07869.pdf > [Consulté le 20/11/2022]
- Tinn, R., Cheng, H., Gu, Y., Usuyama, N., et al. Fine-Tuning Large Neural Language Models for Biomedical Natural Language Processing. arxiv.org, 2021. Disponible à l'adresse : < https://arxiv.org/pdf/2112.07869.pdf > [Consulté le 20/11/2022]
- Thomas, K., Paul, R., Kanzawa, M. PubMeSH : Extreme Multi-label Classification of Biomedical Research. Standford, 2018. Disponible à l'adresse : <https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/custom/15845084.pdf > [Consulté le 20/11/2022]

CINQUIEME PARTIE : MISE AU POINT SUR 2 PROBLEMATIQUES TECHNIQUES

- Copara, j., Knafou,J., Naderi, N., et al. Contextualized French Language Models for Biomedical Named Entity Recognition. In Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier Défi Fouille de Textes. ATALA et AFCP, Nancy, 2022. Disponible à l'adresse : < https://aclanthology.org/2020.jeptalnrecital-deft.4.pdf > [Consulté le 20/11/22]
- Dura, B., Jean, C., Tannier, X., et al. Learning structures of the French clinical language: development and validation of word embedding models using 21 million clinical reports from electronic health records. arxiv.org, 2022. Disponible à l'adresse : https://arxiv.org/pdf/2207.12940.pdf [Consulté le 20/11/22]
- Perera, N., Dehmer, M., Emmert-Streib, F. Named Entity Recognition and Relation Detection for Biomedical Information Extraction. Frontiers in Cell and Developmental Biology [en ligne], 2020, vol.8. Disponible à l'adresse : <https://www.frontiersin.org/articles/10.3389/fcell.2020.00673> [Consulté le 20/11/2020]

- Gérardin, C., Vaillant, P., Wajsbürt, P., et al. Classification multilabel de concepts médicaux pour l'identification du profil clinique du patient. *In* Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT), p.21–30, ATALA : Lille, 2021 Disponible à l'adresse : < Disponible à l'adresse : < https://aclanthology.org/2021.jeptalnrecital-deft.3.pdf> [Consulté le 20/11/2022]
- Islamaj, R., Kwon, D. Kim, S., et al. TeamTat: a collaborative text annotation tool. Nucleic Acids Research, 2020
- Le Clercq de Lannoy, T., Besançon, R., Ferret O., et al. Stratégies d'adaptation pour la reconnaissance d'entités médicales en français. Traitement Automatique des Langues Naturelles (TALN 2022). Avignon, 2022. Disponible à l'adresse : < https://hal.archives-ouvertes.fr/hal-03701500/document> Consulté le [20/11/2022]

SIXEME PARTIE : SOLUTIONS POSSIBLES

- QWAM. INTELLIGENCE ARTIFICIELLE & VALORISATION DES DONNÉES TEXTUELLES. Livre Blanc, 2020.

POUR ALLER PLUS LOIN : SELECTION DE RESSOURCES

Notice : Bibliographie commentée de ressources que j'ai jugé de haute qualité pour approfondir le sujet du TAL et de l'intelligence artificielle. Mis à part : deux références sur le biomédical, toutes les autres références permettent d'aborder le sujet avec un angle plus large. Les références bibliographiques sont classées du général au particulier, puis par ordre d'importance. J'ai utilisé la norme ISO 690.

1 POUR AVOIR UNE VISION ECLAIRE SUR L'INTELLIGENCE ARTIFICIELLE ET LE TAL EN GENERAL

I2D - Information, données & documents 2022/1 (n° 1) : L'INTELLIGENCE ARTIFICIELLE. A.D.B.S. Trimestriel.114p. ISSN en ligne : 2431-3467

« Présente à toutes les étapes du cycle de l'information, l'IA fait évoluer les pratiques documentaires et lance de nouveaux défis pour gérer et valoriser l'information. Dans un environnement en pleine évolution, il s'agit ici de démystifier le sujet de l'IA et d'éclairer les acteurs du champ de l'information professionnelle et scientifique : ce numéro spécial IA offre des repères technologiques et juridiques, présente quelques usages matures et impacts réels de l'IA dans trois domaines clés du secteur de l'info-doc, et donne à réfléchir sur l'évolution des compétences et des métiers qui y sont adossés. » Ghislaine Chartron-Présentation du dossier

Gu, Y., Tinn, R., Cheng, H., Lucas,M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon H. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Trans. Comput. Healthcare[en ligne], 2021, n°1. 23 p. [Consulté le 20/11/2022]. Disponible à l'adresse <https://arxiv.org/pdf/2007.15779.pdf>

Cet article est d'une aide précieuse pour comprendre comment le NLP biomédical diffère du NLP traditionnel. Pour comprendre les bien meilleures performances d'un modèle pré-entraîné sur un domaine spécifique. Pour comprendre le modèle de pointe, dans le domaine biomédical : PubMedBERT. Enfin, pour comprendre la référence, complète, pour les tâches de NLP biomédicales : le benchmark BLURB. Il a des figures très éclairantes sur les mécanismes à l'œuvre dans cette problématique.

Perera, N., Dehmer, M., Emmert-Streib, F. Named Entity Recognition and Relation Detection for Biomedical Information Extraction. Frontiers in Cell and Developmental Biology [en ligne], 2020, vol.8. [Consulté le 20/11/2020]. Disponible à l'adresse : <https://www.frontiersin.org/articles/10.3389/fcell.2020.00673>

L'exploration de texte biomédical et le traitement du langage naturel (BioNLP) est un domaine de recherche intéressant qui traite du traitement des données de revues, de dossiers médicaux et d'autres documents biomédicaux. Compte tenu de la disponibilité de la littérature biomédicale, il y a eu un intérêt croissant pour l'extraction d'informations, de relations et d'idées à partir de données textuelles. Cependant, l'organisation non structurée et la complexité du domaine des documents biomédicaux rendent ces tâches difficiles. Heureusement, certains modèles spécialisés et certaines techniques permettent d'adresser ce problème. Cet article est une plongée dans la reconnaissance d'entité nommées et l'extraction de relations appliqué au domaine biomédical.

2 DEUX BON MANUELS, SUR LE *DEEP LEARNING,* AUX DEMARCHES COMPLEMENTAIRES

Howard, J., Gugger, S. Deep Learning for Coders with Fastai and PyTorch. O'Reilly Media, 2020. 350 p. ISBN : 9781492045526

Les auteurs Jeremy Howard et Sylvain Gugguer montre dans ce manuel : comment entraîner des modèles sur toute une variété de tâches. Ils nous accompagnent pas à pas vers une compréhension en profondeur de la théorie qui sous-tend le *Deep Learning*. L'intérêt de leurs approche : c'est qu'elle fait fi des mathématiques qui dans bon nombre d'ouvrages constituent une véritable barrière à l'entrée, difficile à franchir si on n'a pas les prérequis nécessaires au départ. Dans cet ouvrage, ils démontrent qu'on peut faire tourner des modèles à l'état de l'art avec peu de connaissances en mathématique, peu de données et seulement le minimum requis de code. Enfin, ils nous aident à implémenter les modèles de *Deep Learning* les plus fréquents et nous permettent d'avoir une compréhension entière des algorithmes qui y opèrent en arrière-plan. Dans le détail, ce livre couvre : l'entraînement de modèles en *Computer Vision*, en NLP, sur les données tabulaires, et pour le *collabartive filtering*, l'apprentissage des dernières techniques à la pointe, celles qui sont les plus fonctionnelles à l'usage. Ce manuel nous permet d'améliorer les résultats des modèles en comprenant comment ils marchent, nous apprend à transformer ces modèles en des applications web, à implémenter ces modèles à partir de zéro, tout en nous instruisant sur les implications éthiques de leurs mise en œuvre. Dernier point, ce livre comprend des commentaires d'un des cofondateurs de la librairie très utilisée : PyTorch.

Goodfellow, I., Bengio, Y., Courville A. Deep Learning. MIT Press, 2016. 800 p. ISBN : 9780262035613

Une introduction à un large éventail de sujets liés au *Deep Learning*, couvrant : les connaissances mathématiques et conceptuelles, les techniques d'apprentissage profond utilisées dans l'industrie ainsi que les perspectives de recherche. Le texte offre les connaissances nécessaires mathématiques et conceptuelles, couvrant les concepts pertinents de l'algèbre linéaire, de la théorie des probabilités et de la théorie de l'information, du calcul numérique et de l'apprentissage automatique. Il décrit les techniques d'apprentissage profond utilisées par les praticiens de l'industrie, c'est-à-dire les réseaux de neurones classiques, la régularisation, les algorithmes d'optimisation, les réseaux convolutifs, la modélisation de séquences et la

méthodologie pratique. Il étudie des applications telles que le traitement du langage naturel, la reconnaissance vocale, la vision par ordinateur, les systèmes de recommandation en ligne, la bio-informatique et les jeux vidéo. Enfin, le livre offre des perspectives de recherche, couvrant des sujets théoriques tels que les modèles à facteurs linéaires, les auto-encodeurs, l'apprentissage des représentations, les modèles probabilistes structurés, les méthodes de Monte Carlo, la fonction de partition, l'inférence approximative et les modèles génératifs profonds.

3 DEUX VISIONS ECLAIRANTES SUR LE FONCTIONNEMENT D'UN RESEAU DE NEURONES DU POINT DE VUE CONCEPTUEL ET MATHEMATIQUE

3Blue1Brown. But what is a neural network ? Série de 3 ép. [Vidéos en ligne]. 2017.~20 min [Consulté le 20/11/2022]. Disponible à l'adresse : <https://www.youtube.com/watch?v=aircAruvnKk>

Cette série de vidéos montre comment marche un réseau de neurones de manière imagée et du point de vue des mathématiques qui le sous-tendent. Il le fait étudiant un exemple classique : un réseau de neurones qui apprend à reconnaître les chiffres manuscrits. Il aborde les intuitions derrière l'apprentissage automatique, les intuitions derrière la descente de gradient et l'algorithme de rétropropagation, puis sa traduction mathématique avec le rôle notamment de la « *chain rule* ».

Welch Labs. Neural networks demystified. Série de 7 ép. [Vidéos en ligne].2014.~5 min [Consulté le 20/11/2022]. Disponible à l'adresse : < https://youtu.be/bxe2T-V8XRs >

Dans cette courte série, le vidéaste explicite de manière limpide comment construire et entraîner un réseau de neurones artificiel, en Python, de A à Z. À travers un exemple très simple : il parcourt tous les *process* : Les données de départ et l'architecture du réseau de neurones, la propagation vers l'avant, la descente de gradient, la rétropropagation, la vérification numérique du gradient, l'entraînement. Enfin, le problème de l'*overfitting*, les tests et la régularisation.

1. **Pour rentrer dans le vif du sujet et comprendre les *Transformers* et BERT :**

VASWANI, A., SHAZEER, N., PARMAR, N., *et al.* Attention is all you need. Advances in neural information processing systems[En ligne], 2017, vol. 30.[Consulté le 20/11/2022]. Disponible à l'adresse : < https://arxiv.org/pdf/1706.03762.pdf >

Ce papier commit par une équipe de Google Brain a été publié pour la première fois en 2017 dans la revue NeurIPS. C'est un papier fondateur dans l'émergence du TAL et ses performances en, apprentissage profond, actuelles. Il a été cité plus de 31000. Dans cet article, les auteurs présentent, pour la première fois, l'architecture des *transformers*. C'est une architecture de modèle qui repose entièrement sur un mécanisme d'attention pour déterminer les dépendances globales entre l'entrée et la sortie. Ce papier montre notamment que cette architecture de *transformer* permet une parallélisation beaucoup plus importante et améliore considérablement

la qualité de la traduction après avoir été entraîné pendant à peine douze heures sur huit GPU P100. Le papier se lit bien et est facile à suivre.

Sarkar, A. All you need to know about 'Attention' and 'Transformers' — In-depth Understanding. 2 parties. Towards Data Science, 2022. [Consulté le 22/11/2022] Disponible à l'adresse : <https://towardsdatascience.com/all-you-need-to-know-about-attention-and-transformers-in-depth-understanding-part-1-552f0b41d021>

Il s'agit d'un long article qui parle de presque tout ce qu'il faut savoir sur le mécanisme d'attention, y compris : *Self-Attention, Query, Keys, Values, Multi-Head Attention, Masked-Multi Head Attention*, et *Transformers*, avec en plus quelques détails sur BERT et GPT. L'article est en réalité divisé en deux parties (deux articles consécutifs : part I et part II). Dans la première partie l'auteur couvre tous les blocs d'Attention. Et dans la partie suivante, il expose en profondeur : l'architecture *Transformer*. L'article est complet et très pédagogique : c'est la plus claire explication du mécanisme d'attention que j'ai pu compulser.

Khalid, S. BERT Explained: A Complete Guide with Theory and Tutorial. Medium, 2019. [Consulté le 20/112022]. Disponible à l'adresse : <https://medium.com/@samia.khalid/bert-explained-a-complete-guide-with-theory-and-tutorial-3ac9ebc8fa7c>

Dans la première partie de cet article, l'auteur passe en revue les aspects théoriques de BERT, tandis que dans la deuxième partie, il l'illustre avec un exemple pratique. L'article est construit autour de cinq question : Pourquoi BERT était-il nécessaire ? Quelle est l'idée de base derrière le modèle ? Comment il marche ? Quand peut-on l'utiliser et comment *le finetuner* ? Comment pouvons-nous l'utiliser ? Et enfin, il illustre l'utilisation de BERT pour la classification de texte par un tutoriel.

4 DES POINTS DE DETAILS QUI ONT LEUR IMPORTANCE

HEIDELBERG. Positional embeddings in transformers EXPLAINED | Demystifying positional encodings. [Vidéo en ligne]. 2021. 9min 39. [Consulté le 20/11/2022]. Disponible à l'adresse :<https://www.youtube.com/watch?v=1biZfFLPRSY>

Cette vidéo adresse de manière clair et explicite un des points technique, mais clé, du papier « *Attention is all your need* » : ce que sont les *positional embeddings*, ce qu'ils requièrent. Et elle explique de manière accessible le rôle des fonctions sinus et cosinus dans leurs implémentations.

Khanna, C. Word, Subword, and Character-Based Tokenization: Know the Difference. Towards Data Science, 2021. [Consulté le 20/11/2022]. Disponible à l'adresse : <https://towardsdatascience.com/word-subword-and-character-based-tokenization-know-the-difference-ea0976b64e17>

Dans cet article l'auteur aborde le problème de la *tokenization* et les différents algorithme qui permettent de le traiter. Il commence par définir la *tokenization*, puis il adresse la *tokenization* des mots, puis la *tokenization* basée sur les caractères enfin il adresse celle basée sur les sous-mots. L'article est accessible à un non informaticien.