Industrialisation d'algorithmes de *deep learning* pour l'extraction des caractéristiques des médicaments

S. Bento Pereira[1], R. Benassi[2], Y. Isaac[1], P. Sendorek[2], S. El Alami[2], R. Sagean[2], S. Lequeux[2], N. Cauvet[1]

[1] Vidal SA, 21 rue Camille Desmoulins, 92789 Issy-les-Moulineaux
[2] Publicis Sapient France, 94 avenue Gambetta, 75020 Paris

suzanne.bento-pereira@vidal.fr
romain.benassi@publicissapient.com

Résumé

Les systèmes d'aide à la décision pour le bon usage du médicament nécessitent de disposer d'informations structurées et normalisées pour décrire les caractéristiques des médicaments (indications, contre-indications, effets indésirables etc.). Par le passé ces données étaient saisies manuellement par les pharmaciens chez VIDAL, mais ce processus a été semi-automatisé ces deux dernières années pour faciliter et rendre plus rapide cette tâche. Nous avons implémenté en production un outil d'indexation semi-automatisé dont nous présenterons ici les performances.

Mots-clés

Traitement du langage naturel, Apprentissage automatique, Santé, Médicaments, Résumé des caractéristiques du produit, Terminologie comme sujet.

Abstract

Decision support modules for the proper use of drugs require structured and standardized information to describe the characteristics of drugs (indications, contraindications, adverse effects, etc.). In the past, this data was entered manually by pharmacists at VIDAL, but this process has been semi-automatized over the past two years to make this task easier and faster. We have implemented in production a semi-automatic indexing tool, its performances will be presented here.

Keywords

Natural language processing, Machine learning, Health, Drugs, Summary of product characteristics, Terminology as topic.

1 Introduction

Les modules d'aide à la décision utilisés dans les LAP (Logiciels d'Aide à la Prescription) permettent de sécuriser la prescription des médecins. Ils sont capables de détecter des anomalies qui peuvent mettre en danger les patients comme des contre-indications, des interactions médicamenteuses, des précautions d'emploi etc. Pour pouvoir fonctionner, et être conformes à la réglementation [1], ils doivent disposer au sein de leur base de connaissance des données structurées nécessaires sur tous les produits médicamenteux disponibles sur le marché comme la liste des concepts de contre-indications, d'indications, ou celle des effets indésirables pour chaque médicament.

1.1 Constitution de la base de connaissance sur les médicaments

Pour constituer cette base de données, les pharmaciens VIDAL doivent s'appuyer sur les textes officiels qui contiennent ces informations : les RCP (Résumé des Caractéristiques des Produits). De ces RCP sont extraites les données administratives sur les médicaments telles que le nom du produit, la date de commercialisation etc. et les données thérapeutiques :

- indications
- contre-indications
- précautions d'emploi
- effets indésirables
- etc.

Il existe dans la base de connaissance plus d'une 50^{ne} de type de données différents à renseigner pour un médicament. Et plus de $15\,000$ médicaments sur le marché pour lesquels ces données doivent être régulièrement mises à jour.

1.2 Indexation manuelle des caractéristiques des médicaments à partir des textes officiels

Cette analyse se fait de manière quotidienne, par la lecture des RCP reçus. Chaque RCP (voir Figure 1 pour un exemple) comprend plusieurs rubriques distinctes (ne sont citées que celles qui nous intéressent ici, il en existe une 30^{ne} en tout) :

- la rubrique *Indications thérapeutiques* : narre les maladies pour lesquelles le médicament peut être utilisé

RÉSUMÉ DES CARACTÉRISTIQUES DU PRODUIT

ANSM - Mis à jour le : 11/04/2011

1. DENOMINATION DU MEDICAMENT

DOLIPRANE 1000 mg, comprimé

2. COMPOSITION QUALITATIVE ET QUANTITATIVE

Paracétamol ..... 1000,00 mg

Pour un comprimé.

Pour la liste complète des excipients, voir rubrique 6.1.

3. FORME PHARMACEUTIQUE

Comprimé.

4. DONNEES CLINIQUES

4.1. Indications thérapeutiques

Traitement symptomatique des douleurs d'intensité légère à modérée et/ou des états fébriles.
Traitement symptomatique des douleurs de l'arthrose.

4.2. Posologie et mode d'administration

Mode d'administration
Voie orale.
Les comprimés sont à avaler tels quels avec une boisson (par exemple eau, lait, jus de fruit).
Posologie
Attention: cette présentation contient 1000 mg de paracétamol par unité: ne pas prendre 2 unités à la fois.
Cette présentation est réservée à l'adulte et à l'enfant à partir de 50 kg (environ 15 ans).
La posologie unitaire usuelle est de un comprimé à 1000 mg par prise, à renouveler au bout de 6 à 8 heures. En cas de besoin, la prise peut être répétée au bout de 4 heures minimum.
Il n'est généralement pas nécessaire de dépasser 3 g de paracétamol par jour, soit 3 comprimés par jour.
Cependant, en cas de douleurs plus intenses, la posologie maximale peut être augmentée jusqu'à 4 g (4 comprimés) par jour. Toujours respecter un intervalle de 4 heures entre deux prises.
Fréquence d'administration:
Les prises systématiques permettent d'éviter les oscillations de douleur ou de fièvre:

- chez l'adulte, elles doivent être espacées de 4 heures minimum.

Insuffisance rénale:
En cas d'insuffisance rénale sévère (clairance de la créatinine inférieure à 10 ml/min), l'intervalle entre deux prises sera au minimum de 8 heures.
Ne pas dépasser 3 g de paracétamol par jour, soit 3 comprimés.

4.3. Contre-indications

- Hypersensibilité au paracétamol ou aux autres constituants.
- Insuffisance hépatocellulaire.

4.4. Mises en garde spéciales et précautions d'emploi

Figure 1: Le RCP du DOLIPRANE 1 000 mg cp (`http://agence-prd.ansm.sante.fr`).

- la rubrique *Contre-indications* : décrit les situations dans lesquelles la prise du médicament est dangereuse
- la rubrique *Effets indésirables* : explicite les effets non souhaités, secondaires au traitement par le médicament et aboutissant à un résultat néfaste (gêne, allergie, complications graves, y compris le décès)

Après lecture et analyse, vient la saisie des données une à une suivant les terminologies existantes et gérées par ailleurs chez VIDAL (voir 3.1.1). Les pharmaciens vont choisir parmi les concepts disponibles de chaque terminologie ceux qui correspondent le mieux à la notion qu'ils souhaitent indexer. Cette saisie se fait via des formulaires dans des applications internes. En moyenne, les pharmaciens vont renseigner 44 termes pour les effets indésirables, 8 pour les contre-indications et 4 termes pour les indications d'un médicament ce qui est un travail assez fastidieux.

1.3 Faciliter l'indexation via l'IA

L'*apprentissage automatique*, ou *Machine Learning* (ML), nous permet de semi-automatiser cette tâche pour la rendre plus rapide et plus simple pour les pharmaciens. Nous avons implémenté en production un outil d'indexation semi-automatisé qui suggère les concepts potentiellement pertinents et positionnés sur les phrases du RCP. Son fonctionnement intègre aujourd'hui l'indexation des indications, contre-indications et effets indésirables. Sur la base de ces propositions, le pharmacien peut décider de valider, supprimer ou modifier les concepts proposés afin de renseigner ces données dans la base de connaissance. Nous présenterons dans cet article le fonctionnement de notre outil basé sur deux approches ML et une approche à base de règles. Nous présenterons également ses performances. Et nous terminerons par une discussion et conclusion.

2 État de l'art

L'indexation de concepts dans les documents consiste à y détecter la présence de concepts d'un référentiel termino-ontologique. Cela permet de rendre l'information qui était jusque-là inexploitable sous forme de texte brut, exploitable par des applications informatiques. Par exemple, dans le domaine médical, les concepts MeSH sont employés pour l'indexation et la recherche d'articles scientifiques dans la base MEDLINE[1], et ceux de la Classification Internationale des Maladies (CIM10) sont employés pour caractériser les séjours hospitaliers à des fins médico-économiques[2].

L'indexation automatique de concepts terminologiques dans des documents de santé a été abordée dans de nombreux travaux de recherches avec l'utilisation de plusieurs techniques. Il existe des approches à base de TAL (Traitement Automatique des Langues) avec l'utilisation de *regex*, dictionnaires et terminologies existantes [2]. Ce sont des approches qui sont assez efficaces lorsqu'elles s'appliquent à des terminologies riches en termes et synonymes et quand les termes sont proches de ceux que nous trouverons dans le texte à analyser. Ces approches ont été par exemple déjà utilisées pour l'extraction de concepts français dans les RCP [3]. Elles ont aussi le mérite de ne pas nécessiter de bases d'apprentissage importantes qui sont souvent inexistantes ou non disponibles.

Des méthodes à base de ML peuvent aussi être utilisées avec des classifieurs type *conditional random fields* (CRFs), *support vector machines* (SVM), *Convolutional Neural Network* (CNN) ou *Transformers* (BERT). Les résultats sont très bons lorsque ces méthodes sont appliquées sur des corpus parfaitement annotés et des terminologies assez restreintes [4, 5, 6, 7, 8, 9]. En particulier, Rubrichi et al. [4] qui a comparé deux approches à base de classifieurs CRFs et SVM pour l'indexation des interactions médicamenteuses en italien à partir des RCP.

Enfin il y a des approches hybrides mêlant *apprentissage automatique* et TAL. Par exemple, Zweigenbaum et al. [10] pour l'indexation de concepts CIM10 en français dans les certificats de décès utilise une méthode à base de dictionnaires puis des classifieurs SVM.

La plupart des études sont toutefois appliquées à la langue anglaise, le français l'est beaucoup moins. Et nous avons trouvé très peu d'articles sur l'indexation de RCP.

3 Méthodes

3.1 La base de documents et les données disponibles

Les données ont été récupérées à partir du système d'information VIDAL. Elles comprennent une base de documents RCP, l'indexation manuelle correspondante et les 3 terminologies d'indexation.

[1] `pubmed.ncbi.nlm.nih.gov`
[2] `epmsi.atih.sante.fr/welcomeEpmsi.do`

3.1.1 Les trois terminologies

Indications. Cette terminologie contient la liste des concepts décrivant les indications. Elle contient $4\,047$ concepts. Chacun d'entre eux a un identifiant unique, un libellé préféré et 0 à n synonymes.

Contre-indications. Elle liste $3\,806$ concepts décrivant les contre-indications. Chacun d'entre eux a également un identifiant unique, un libellé préféré et 0 à n synonymes.

Effets indésirables. Contient $4\,714$ concepts pour les effets indésirables avec pour chacun un identifiant unique, un libellé préféré et 0 à n synonymes.

3.1.2 La base de documents indexés manuellement

Nous disposons au début du projet d'une base de $9\,353$ RCP indexés, issus de 19 années d'historique d'indexation manuelle par les pharmaciens VIDAL.
L'indexation manuelle est réalisée au niveau du document, elle lie l'identifiant du document avec les identifiants des concepts indexés pour les indications, contre-indications et effets indésirables (voir un exemple dans le Tableau 1). Il n'existe pas de lien entre la rubrique ou la phrase exacte du document et les identifiants des concepts indexés.

Type de données\Produit	PD111 - DOLIPRANE 1000 mg cp
Indications	IND45 Fièvre; IND89 Douleur d'intensité légère à modérée
Contre-indications	CI02 Hépatopathie décompensée; CI46 Hypersensibilité au paracétamol; CI56 Insuffisance hépatique sévère
Effets indésirables	EI12 Céphalée; EI45 Anémie; EI89 Diarrhée; EI87 Confusion mentale; EI43 Malaise; EI15 Vertige

Table 1: Extrait des indications, contre-indications et effets indésirables renseignés avec leurs identifiants pour le médicament DOLIPRANE $1\,000$ mg cp

La base d'apprentissage contient l'extraction de $8\,353$ rubriques *Indications thérapeutiques*, $8\,353$ rubriques *Contre-indications* et $8\,353$ rubriques *Effets indésirables* indexées. Elle comprend aussi les indexations des concepts contre-indication, indication et effet indésirables correspondants. Elle sera utilisée pour entraîner nos algorithmes d'IA.

3.2 Une combinaison de trois approches

L'objectif est de trouver, parmi une liste pré-définie, l'ensemble des concepts positionnés sur une phrase. Chaque concept est représenté par un label préféré, c'est-à-dire un groupe de mot le caractérisant mais aussi par un ensemble de synonymes, entendus ici comme au sens de groupes de mots potentiellement différents du label préféré mais recouvrant la même réalité.
Nous avons fait le choix de combiner trois types d'approches différentes afin d'optimiser les performances. Deux de ces approches reposent sur une mécanique d'*apprentissage automatique* et sont ensuite combinés dans le cadre d'un fonctionnement dit *hybride*. La dernière approche, quant à elle, se situe en aval des deux premières. Elle s'inscrit dans une logique de rattrapage, à partir de règles métier, de cas bien identifiés par les équipes VIDAL.

3.2.1 Approche par similarité

Distance de Ratcliff-Obershelp. Afin de caractériser la présence d'un concept au sein d'un texte, nous utilisons la mesure de similarité de Ratcliff-Obershelp [11]. Pour deux chaînes de caractères S_1 et S_2, cette mesure se calcule selon la formule suivante

$$D = \frac{2K}{|S_1| + |S_2|} \tag{1}$$

où $|.|$ représente l'opérateur donnant le nombre de caractères d'une chaîne, et K le nombre de *caractères correspondants* entres les deux chaînes. Ce concept de nombre de *caractères correspondants* se définit récursivement par la somme de la taille de la plus grande sous-chaîne en commun et le nombre de *caractères correspondants* des deux côtés de ladite plus grande sous-chaîne en commun.

Dans le cadre de notre application. Afin d'identifier la présence de l'un des concepts d'intérêt au sein du texte donné en entrée, nous allons parcourir l'ensemble des libellés possibles. Pour chaque synonyme nous calculons la mesure de similarité entre chacun des m mots le constituant avec chacun des n mots du texte. Cela revient à construire un tableau de taille $[m, n]$ où la valeur à la $i^{\text{ème}}$ ligne et $j^{\text{ème}}$ colonne correspond à la similarité entre le $i^{\text{ème}}$ mot du synonyme, et le $j^{\text{ème}}$ mot du texte.
Par exemple, si nous recherchons un concept *AVC* représenté par le synonyme *accident vasculaire cérébral* dans un texte qui contiendrait consécutivement chacun des trois mots le constituant, et des mots très différents de part et d'autre, nous aurions un résultat proche du suivant

$$\begin{aligned}
&\ldots 0, 0, 1, 0, 0, 0, 0, 0 \ldots && \text{similarité avec } accident\\
&\ldots 0, 0, 0, 1, 0, 0, 0, 0 \ldots && \text{similarité avec } vasculaire\\
&\ldots 0, 0, 0, 0, 1, 0, 0, 0 \ldots && \text{similarité avec } cérébral.
\end{aligned}$$

Prendre le maximum par colonne nous permet d'aboutir au vecteur suivant

$$\ldots 0, 0, 1, 1, 1, 0, 0, 0 \ldots,$$

nous calculons alors une moyenne glissante sur une fenêtre de taille m, ce qui revient à considérer un calcul de convolution avec un filtre $[1/m, 1/m, \ldots, 1/m]$. Dans le cas de notre exemple, m est égal à 3, nous obtenons alors le filtre $[1/3, 1/3, 1/3]$ et le résultat suivant:

$$\ldots 0, 0, \mathbf{0.3}, \mathbf{0.6}, \mathbf{1}, \mathbf{0.6}, \mathbf{0.3}, 0 \ldots.$$

Le maximum de ce vecteur correspond à ce que nous appelons ici le *maximum de similarité* (et l'indice de ce maximum dans le vecteur peut permettre de récupérer sa position dans le texte).

Il suffit que la valeur de ce *maximum de similarité* soit supérieur à un seuil fixé pour que le concept soit considéré comme présent dans le texte. Nous avons fait le choix d'utiliser une valeur de seuil spécifique par synonyme et d'en faire l'apprentissage automatiquement.

Apprentissage des seuils par modèle. Pour un concept donné, et un synonyme s, l'objectif ici est de déterminer les seuils λ_s, qui vont maximiser la f-mesure associée à la détection du concept parmi un ensemble de textes d'entraînement. Pour chaque synonyme nous considérons un ensemble d'entraînement spécifique contenant tous les textes pour lesquels le concept n'est pas présent, mais seulement un sous-ensemble des textes pour lequel il est présent. Ce sous-ensemble de textes contenant le concept est choisi parmi les textes les plus proches du synonyme, c'est-à-dire dont la valeur de similarité est inférieure à λ_s. Il suffit alors de tester comme valeur de seuil l'ensemble des valeurs de similarité atteintes pour le synonyme donné et vérifier celle qui maximise la f-mesure.

3.2.2 Approche par réseaux convolutifs

Contexte. En complément de l'approche par similarité, et à l'instar de ce que ferait le pharmacien humain, nous avons développé une méthode d'Intelligence Artificielle (IA) dont l'objectif est de s'intéresser directement au sens du texte par opposition à une simple recherche de mots, ou groupe de mots. Nous nous appuyons sur les récentes avancées du *deep learning* dans le domaine du NLP (*Natural Language Processing*) [12, 13, 14, 15, 16, 17] afin d'obtenir un niveau de représentation particulièrement fin du langage.

***Embeddings* et CNN.** Ces méthodes permettent de produire des représentations numériques des mots via une approche nommée *plongement lexical*, ou *word embedding*, dont la spécificité est que deux mots de sens proche y ont une représentation numérique proche elle aussi (voir par exemple [12, 13]). L'état de l'art actuel de ces approches, se fonde sur des architectures de type *Transformers*, introduites dans [15] et dont le modèle BERT [16] est disponible en libre accès. Les RCP traités étant en français c'est logiquement CamemBERT [17], la déclinaison francophone de BERT, qui a été utilisée pour la génération des *embeddings*. Une fois les textes des RCP plongés dans un espace numérique, l'enjeu est toujours d'être capable d'y reconnaître, ou non, la présence des concepts. Plusieurs approches sont possibles (voir notamment [18]) parmi lesquelles nous avons fait le choix d'utiliser une architecture de type CNN par analogie entre la recherche d'un motif dans un texte et celle d'un objet dans une image, cas d'usage classique des CNN ([19, 20, 21]).

Mise en œuvre. La présence d'un concept au sein d'un texte est modélisée comme étant un événement indépendant de la présence respective des autres concepts. En particulier, la présence de l'un n'exclut pas la présence d'un autre. Nous sommes donc en présence d'une classification type *multi-labels* plutôt que *multi-classes*, cette dernière dénomination supposant un caractère exclusif des catégories les unes par rapport aux autres. Cette caractéristique ainsi que la présence d'un grand nombre de concepts à extraire, plusieurs milliers, nous incitent à construire une modélisation spécifique pour *chacun* des concepts.

Le réseau de neurones que nous considérons pour *un concept donné* peut s'écrire sous la forme

$$\hat{p} = \max_t F(c[t-k : t+k]) \qquad (2)$$

où $\hat{p}$ est une probabilité caractérisant la présence du concept au sein du texte, F est une fonction correspondant à une succession de filtres convolutifs, c est la séquence d'*embeddings* de texte issue de CamemBERT, t l'indice sur lequel est centrée l'évaluation et k la taille du filtre de convolution. Nous appelons cette architecture *common denominator* puisqu'elle est conçue afin de reconnaître le dénominateur commun (en termes de sémantique) à chaque concept.

Le réseau contient n couches convolutives successives, chacune ayant un paramètre de dilatation égal à 2^i pour i son numéro dans l'ordre de succession. Le nombre n est choisi de manière à être pertinent par rapport à la taille des libellés. Le réseau (voir Figure 2) se compose ainsi de n couches de convolution avec pour chacune une fonction d'activation tangente hyperbolique, sauf pour la dernière qui se voit affectée une fonction sigmoïde afin d'obtenir des valeurs entre 0 et 1, caractérisant une probabilité de présence du concept. Le réseau se termine par une couche de *max pooling* finale qui permet de ne récupérer qu'une valeur de probabilité, la maximale caractérisant à elle seule la présence ou l'absence du concept sur l'ensemble du texte d'intérêt. Afin de construire le jeu d'apprentissage pour chacun des modèles, nous associons à chaque texte une valeur cible (ou *target*) valant 0 ou 1 en fonction respectivement de l'absence ou présence du concept (voir Figure 3). À noter qu'ici, contrairement à l'approche par similarité décrite en 3.2.1, nous n'utilisons que les textes bruts sans aucune information de synonymie. Charge au réseau d'apprendre à reconnaître le dénominateur commun sémantique d'un concept à partir des seuls textes et des valeurs cibles.

Le texte est ensuite transformé sous forme de vecteurs numériques (ou *embeddings*) grâce au modèle CamemBERT selon une procédure en deux temps. La première étape consiste à séparer le texte en *tokens*, des éléments plus petits mais porteurs d'unité de sens pour le modèle (ce ne sont pas nécessairement des mots, ce peut-être seulement des groupes de lettres successives). La seconde revient à associer à chaque *token* un vecteur numérique de taille M (ici, M vaut 768). Le modèle CamemBERT est pré-entraîné pour effectuer automatiquement ces deux tâches, c'est donc ainsi que nous l'utilisons pour transformer nos données d'entrée. Prenons l'exemple d'un texte réduit à une phrase

texte: *Ce médicament ne doit jamais être utilisé dans les états de rétention hydrosodée*,

la séparation en *tokens* nous donne ici vingt éléments sous la forme d'une série d'identifiants,

$$\mathbf{tokens} : \big(\ 5 \quad 44 \quad 31922 \quad 45 \quad 279 \quad 283 \quad \dots \quad 6\ \big).$$

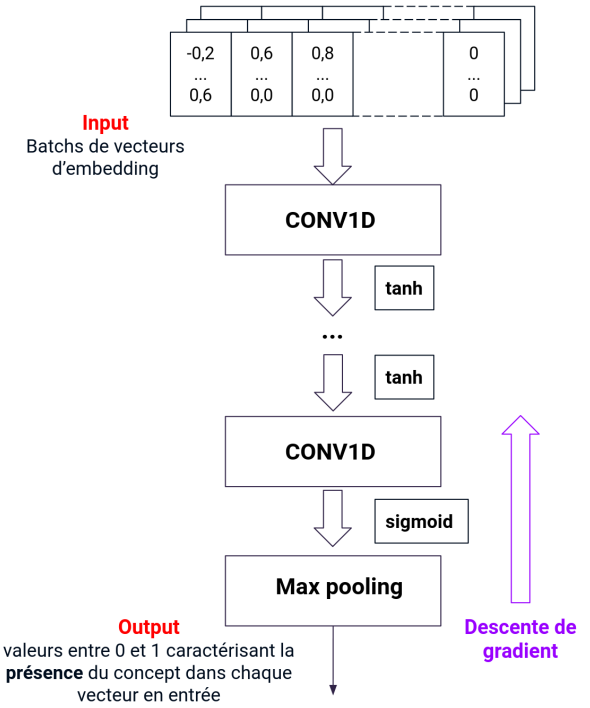

Figure 2: Architecture du réseau *common denonimator* conçu dans le cadre de ce projet.

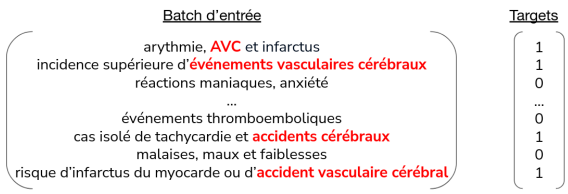

Figure 3: Illustration de la construction d'un jeu d'apprentissage pour un concept donné (ici *accident vasculaire cérébral*).

Chacun de ces *tokens* est ensuite plongé dans un espace vectoriel numérique de dimension 768 ce qui permet d'obtenir en sortie, sur cet exemple, une matrice de taille 768×20,

$$\textbf{embeddings} : \begin{pmatrix} -0.2 & 0.6 & \dots & -0.4 \\ 0.7 & -0.1 & \dots & -0.8 \\ \vdots & \vdots & \vdots & \vdots \\ 0.6 & 0.0 & \dots & -0.4 \end{pmatrix}.$$

Pour entraîner le réseau, nous cherchons à optimiser une fonction de coût qui correspond, à nouveau, à la f-mesure. Concrètement cela revient à optimiser la fonction suivante

$$-\log\left[f_m(\hat{\text{p}}, \text{targets})\right] \tag{3}$$

où f_m correspond à une version dite *soft* de la f-mesure, variante continue de la version traditionnelle, permettant d'aider la convergence lors de l'apprentissage.

3.2.3 Rattrapage à base de règles métiers

Les deux premières approches sont complétées par une troisième, qui va selon plusieurs règles métiers ajuster les propositions de concepts en en ajoutant ou en supprimant.
Ces règles associent une portion de texte provenant de RCP à un ou plusieurs concepts provenant des trois terminologies avec une indication *à ajouter* ou *à supprimer*.
Ces règles ont été créées à base de règles métiers (après discussions avec les pharmaciens) et à base de motifs en erreurs trouvés après plusieurs tests sur les approches IA. Il existe plus d'un millier de règles.
Un exemple de règle : *Infections et infestations* qui est un titre dans les tableaux de la rubrique *Effets indésirables* ne doivent pas être indexés avec le concept EI79 *infection*.

3.2.4 Combinaison des trois approches

Les trois approches considérées, pour rappel,

1. similarité
2. *common denominator*
3. rattrapage à base de règles métiers,

se combinent et se complètent. Les deux premières relèvent d'un apprentissage automatique et, comme vu précédemment, pour chaque concept deux modèles sont appris. L'un pour l'approche par similarité, l'autre via entraînement d'un réseau de neurones convolutif selon le principe de recherche du dénominateur commun. Dans les deux cas un calcul de la f-mesure est effectué, sur un jeu de données de test, afin d'évaluer les performances respectives des deux modèles. Les f-mesures sont sauvegardées en base, modèle par modèle (voir Figure 4).
Une fois en production, l'annotation des RCP se déroule en deux étapes. Dans un premier temps, un système *hybride* entre l'approche de similarité et l'approche *common denominator* se met en place autour de la table des f-mesures enregistrées durant l'apprentissage. Pour chaque concept d'intérêt c'est, parmi les deux disponibles, le modèle ayant eu la f-mesure la plus favorable sur le jeu de

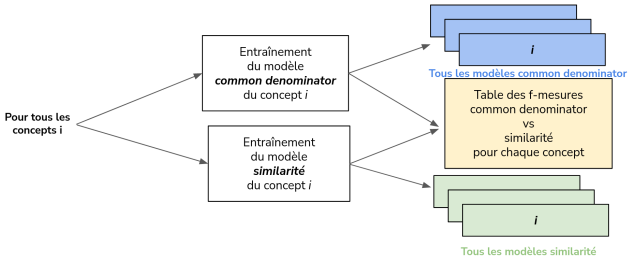

Figure 4: *Entraînement* - pour chaque concept, les modèles *similarité* et *common denominator* sont entraînés. Les valeurs respectives des f-mesures des deux modèles sont gardées en base de données afin de pouvoir décider ultérieurement lors de l'annotation lequel des deux modèles choisir concept par concept.

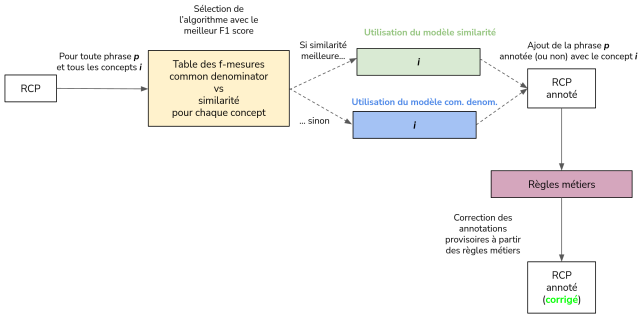

Figure 5: *Annotation* - pour chaque concept c'est, parmi les deux modèles statistiques entraînés, celui ayant eu la f-mesure la plus favorable qui est choisi pour annoter le RCP puis une correction à base de règles métiers déterministes est effectuée.

test qui est choisi. À l'issue de cette première itération, une version annotée *intermédiaire* du document est obtenue ; celle-ci sert alors de base à une seconde étape, le rattapage par règles métiers, qui permet d'amender le document et d'en faire une version *corrigée* (voir Figure 5), à valider par l'humain.

3.3 Intégration dans l'outil d'indexation semi-automatique

Le service de suggestion de concepts permettant l'indexation semi-automatique du RCP est disponible via une API qui prend en entrée le texte HTML de la rubrique et le nom de la rubrique concernée. En sortie, elle délivre le contenu de la rubrique découpée en phrases avec les indexations trouvées automatiquement selon la terminologie adéquate.

Cette API a été intégrée au sein de l'outil semi-automatique d'indexation qui présente le texte du RCP ainsi que les propositions de concepts faites par l'IA à valider par les pharmaciens. Pour le moment l'application ne propose que les concepts de contre-indication et d'effets indésirables (les indications seront intégrées dans les prochains mois).

Les propositions apparaissent sur les phrases encadrées de rouge et après passage de la souris. Et chaque proposition

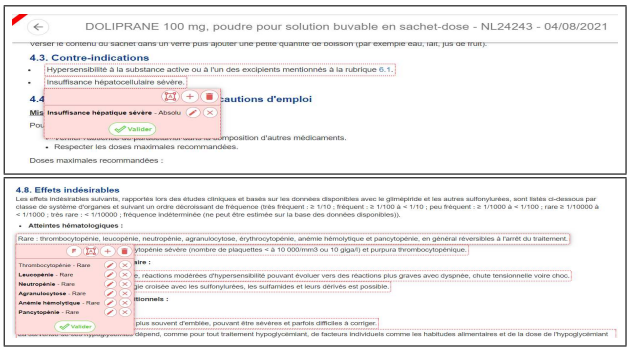

Figure 6: Copies d'écran de l'outil d'indexation semi-automatique (pour la rubrique *Contre-indications* en haut et la rubrique *Effets indésirables* en bas).

dispose d'une possibilité de suppression, modification et de validation (voir Figure 6).

L'indexeur pharmacien va analyser l'intégralité des deux rubriques contre-indication et effets indésirables et les propositions associées afin d'annoter l'ensemble des concepts nécessaires pour ce document.

3.4 La Boucle de *Feedback*

À chaque RCP analysé et son indexation complète validée, un document contenant le RCP indexé est enregistré dans les bases documentaires VIDAL.

Ces documents indexés sont récupérés tous les mois pour alimenter la base d'apprentissage et bénéficier de nouvelles entrées plus récentes. Les référentiels sont aussi mis à jour au même moment, et une nouvelle version de l'API est déployée.

Ceci nous permet de pouvoir apprendre en continu sur les nouvelles indexations.

3.5 Analyse des performances

Nous avons mesuré les performances de l'API d'indexation automatique.

La base d'évaluation contient 1 000 rubriques *Indications thérapeutiques*, 1 000 rubriques *Contre-indications* et 1 000 rubriques *Effets indésirables* indexées soit environ 10% du total de départ. Nous disposons comme *gold standard* des indexations manuelles correspondantes réalisées par les pharmaciens avec les concepts d'indication, de contre-indication et d'effets indésirables. Elle est utilisée pour évaluer la qualité de l'apprentissage.

L'API a été utilisée sur les textes des rubriques et nous avons comparé les résultats d'indexation automatique obtenus avec le *gold standard*. Les différentes mesures calculées sont :

- La précision : proportion de concepts pertinents parmi l'ensemble des concepts suggérés automatiquement par l'outil (niveau document)

$$\frac{|C_{manuel} \cap C_{auto}|}{|C_{auto}|} \quad (4)$$

avec C_{manuel} l'ensemble des concepts pertinents, C_{auto} celui des concepts obtenus via l'outil et $|.|$ l'opérateur donnant le cardinal d'un ensemble.

- Le rappel : proportion de concepts pertinents suggérés automatiquement parmi l'ensemble des concepts manuellement indexés

$$\frac{|C_{manuel} \cap C_{auto}|}{|C_{manuel}|} \quad (5)$$

- La f-mesure : moyenne harmonique de la précision à plat et du rappel

$$2\,\frac{\text{precision . rappel}}{\text{precision} + \text{rappel}} \quad (6)$$

Les mesures seront également détaillées par terminologie.

4 Résultats des performances de l'indexation automatique

	Nombre	Précision	Rappel	f-mesure
Indications	1 000	0.89	0.89	0.87
Contre-indications	1 000	0.92	0.87	0.88
Effets indésirables	1 000	0.81	0.86	0.83
TOTAL	1 000	0.87	0.87	0.86

Table 2: Résultats sur la base d'évaluation

Les évaluations sur les 1 000 rubriques indexées pour chaque rubrique *Indications*, *Contre-indications* et *Effets indésirables* a montré que l'IA obtient une performance moyenne totale de 86% de f-mesure (voir Tableau 2). La précision moyenne mesurée étant de 87% et le rappel moyen de 87%[3]. Le détail par rubrique, montre que les meilleurs résultats sont obtenus pour les *Contre-indications* avec 88% de f-mesure moyenne soit 1% de plus que pour les *Indications* et 5% de plus pour les *Effets indésirables*.

5 Discussion

Les résultats obtenus sont satisfaisants, l'IA permet d'extraire beaucoup de concepts pertinents et de faire gagner du temps aux utilisateurs dans la recherche de ces concepts.
L'avis des utilisateurs après utilisation en production et au quotidien du nouvel outil d'indexation semi-automatique est bon également.
Il reste encore toutefois une marge de progression pour cet outil. Afin de comprendre ces résultats, les erreurs ont été analysées afin de lister les causes et dégager des pistes d'amélioration. Ces analyses ont été menées

[3]Les valeurs de f-mesures du Tableau 2 ne correspondent pas directement à la moyenne harmonique des valeurs de précisions et de rappels indiquées, mais à la moyenne des f-mesures obtenues sur l'ensemble des rubriques. Numériquement, ceci implique que la f-mesure n'est pas nécessairement comprise entre les valeurs correspondantes de précision et de rappel moyens.

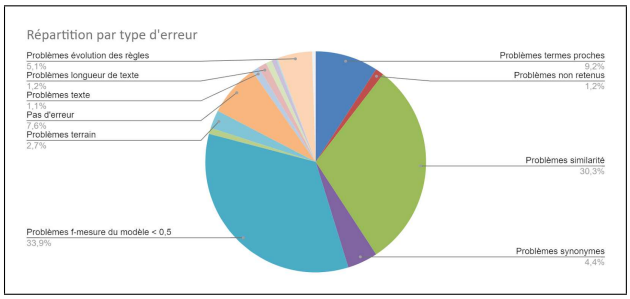

Figure 7: Répartition des erreurs sur les EI.

avec deux pharmaciens indexeurs experts sur l'indexation automatique et sur deux échantillons de rubriques :

- Un échantillon de 15 rubriques *Effets Indésirables*. Elles ont été choisies selon les critères suivants : une f-mesure globale d'environ 60%, une rubrique assez longue avec de nombreuses phrases et de nombreux concepts indexés. C'est ainsi 555 concepts en erreur (soit des concepts manquants soit des concepts erronés) qui ont été analysés.

- Et un échantillon de 17 rubriques *Contre-indications* choisies selon les mêmes critères. Ici c'est 171 concepts en erreur qui ont été analysés.

Concernant l'indexation des effets indésirables, plusieurs causes ont été identifiées (voir Figure 7) :

- problème de performance de certains modèles : leur performance propre ne dépasse pas 50% de f-mesure. Leurs sous-performances peuvent s'expliquer par différents facteurs liés à la base d'apprentissage. D'abord, il existe pour certains concepts peu d'exemples avec seulement une ou deux indexations, c'est le cas de concepts rares ou de concepts récemment ajoutés (voire pas d'indexation du tout pour les nouvelles notions). Ensuite certains termes sont ambigus, mêmes libellés ou synonymes alors que le sens est différent, ils ont tendance à sortir en même temps ou l'un à la place de l'autre (exemple : *mg* qui correspond à une unité de mesure et non au concept *magnésium*). Enfin il existe des difficultés pour les termes proches (*asthme/asthme sévère*). Les phrases et contextes étant souvent proches, il est difficile pour les modèles associés d'être performant

- problème de longueur de texte : la prédiction d'indexation se fait au niveau des phrases. Cette longueur de texte est parfois insuffisante pour retrouver la notion au complet ou son niveau de précision suffisant

- pas d'erreur : l'indexation manuelle est réalisée par différents indexeurs qui peuvent avoir des avis différents sur le choix du terme le plus pertinent

- les règles d'indexation peuvent aussi évoluer dans le temps : une indexation qui était considérée comme

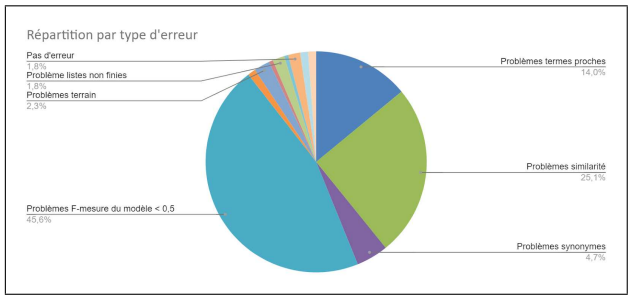

Figure 8: Répartition des erreurs sur les CI.

juste auparavant peut ne plus l'être après changement des règles éditoriales d'indexation par les pharmaciens

- problème de terrain : certains concepts ne sont pas à indexer alors qu'ils sont bien présents dans le texte. C'est le cas des terrains qui décrivent les facteurs favorisant la survenue d'un effet indésirable mais pas l'effet indésirable en lui même (exemple : *risque d'éruption cutanée chez les personnes porteuses du VIH*, seul le *risque d'éruption cutanée* est indexé par les indexeurs, pas le terrain *VIH*)
- problème de texte : la qualité du document, son orthographe, son format peuvent avoir une incidence sur l'indexation automatique
- problème de similarité : la méthode par similarité va avoir tendance à rapprocher des libellés proches (à une ou deux lettres près) ce qui va entraîner l'indexation d'un mauvais terme
- problème de synonymes : les terminologies utilisées ne sont pas exhaustives en matière de synonymes. Il peut manquer certains libellés qui vont empêcher l'indexation d'un concept particulier.
- listes non finies d'éléments : certains ne sont pas explicités dans le texte avec l'utilisation de notions telles que *autres*, *etc.* ou des listes entres parenthèses non finies
- il est également possible, en cas de nécessité, pour l'indexeur d'indexer une propriété clinique absente du RCP mais indiquée par d'autres sources d'information ou de ne pas retenir des termes présents dans le RCP (règles d'indexation particulières).

Concernant l'indexation des contre-indications, les mêmes types de causes peuvent être remontés avec des proportions un peu différentes (voir Figure 8) : La boucle de *Feedback* et l'ajout continu de règles nous permettent d'améliorer petit à petit les propositions automatiques, il sera donc intéressant à terme de faire une analyse de l'évolution des performances.
Pour la suite, plusieurs actions vont être menées. Côté gestion de terminologies, il est prévu d'enrichir les terminologies en synonymes. Côté indexeurs pharmaciens, il est prévu de réaligner les façons d'indexer pour ne plus avoir des soucis de différences d'indexation inter-indexeurs. Côté data science, un algorithme spécifique va être travaillé pour les problèmes de terrain.
Dans le futur, il est envisagé d'intégrer d'autres rubriques, toujours réalisées à la main actuellement par les pharmaciens (la rubrique *Composition qualitative et quantitative* avec l'indexation des substances, la rubrique des *Précautions d'emploi* avec l'indexation des précautions d'emploi).

6 Conclusion

L'objectif du projet était de développer des outils d'aide à l'indexation permettant de suggérer automatiquement des concepts à indexer aux pharmaciens pour l'indexation des propriétés thérapeutiques des médicaments dans la base de connaissances VIDAL. Trois méthodes ont été mises en œuvre : deux approches utilisant des algorithmes d'apprentissage automatique et une utilisant directement des règles métiers.
L'étude démontre qu'une automatisation d'une partie de l'indexation est possible à hauteur de 86% pour aider les indexeurs pharmaciens dans l'indexation quotidienne des RCP.

Remerciements

Remerciements aux équipes VIDAL : les pharmaciens, les développeurs et les *data scientists* qui ont participé aux développements de l'outil.

Références

[1] HAS[4]. Certification des logiciels des professionnels de santé. *Mis à jour le 03 mai 2021*.

[2] Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. *O'Reilly Media, Inc.*.

[3] Bento Pereira, S. (2008). *Indexation multi-terminologique de concepts en Santé* (Doctoral dissertation, Rouen).

[4] Rubrichi, S., & Quaglini, S. (2012). Summary of Product Characteristics content extraction for a safe drugs usage. *Journal of Biomedical Informatics*, 45(2), 231-239.

[5] Hasan, M., Kotov, A., Carcone, A. I., Dong, M., Naar, S., & Hartlieb, K. B. (2016). A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories. *Journal of biomedical informatics*, 62, 21-31.

[6] Blanco, A., Perez-de-Viñaspre, O., Pérez, A., & Casillas, A. (2020). Boosting ICD multi-label classification of health records with contextual embeddings and label-granularity. *Computer methods and programs in biomedicine, 188*, 105264.

[4] `https://www.has-sante.fr`

[7] Remmer, S., Lamproudis, A., & Dalianis, H. (2021). Multi-label Diagnosis Classification of Swedish Discharge Summaries–ICD-10 Code Assignment Using KB-BERT. In *RANLP 2021: Recent Advances in Natural Language Processing, 1-3 Sept 2021, Varna, Bulgaria* (pp. 1158-1166). Association for Computational Linguistics.

[8] Amin, S., Neumann, G., Dunfield, K., Vechkaeva, A., Chapman, K. A., & Wixted, M. K. (2019, September). MLT-DFKI at CLEF eHealth 2019: Multi-label Classification of ICD-10 Codes with BERT. In CLEF *(Working Notes)* (pp. 1-15).

[9] Biseda, B., Desai, G., Lin, H., & Philip, A. (2020). Prediction of ICD Codes with Clinical BERT Embeddings and Text Augmentation with Label Balancing using MIMIC-III. *arXiv preprint arXiv:2008.10492*.

[10] Zweigenbaum, P., & Lavergne, T. (2017). Détection de concepts et granularité de l'annotation (Concept detection and annotation granularity). In *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2-Articles courts* (pp. 226-233).

[11] Ratcliff, J.W., & Metzener, D. E. (1988). Pattern Matching: The Gestalt Approach. *Dr. Dobb's Journal*, 13(7), 46.

[12] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

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

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[16] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[17] Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., & Sagot, B. (2020) CamemBERT: a Tasty French Language Model, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Jul, 2020, pages 7203–7219

[18] Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. *arXiv preprint arXiv:1702.01923*.

[19] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.

[20] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

[21] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).