

Désidentification de comptes-rendus hospitaliers dans une base de données OMOP

Nicolas Paris, Matthieu Doutreligne, Adrien Parrot, Xavier Tannier

▶ **To cite this version:**

Nicolas Paris, Matthieu Doutreligne, Adrien Parrot, Xavier Tannier. Désidentification de comptes-rendus hospitaliers dans une base de données OMOP. TALMED 2019 : Symposium satellite francophone sur le traitement automatique des langues dans le domaine biomédical, Aug 2019, Lyon, France. hal-02564721

HAL Id: hal-02564721
https://hal.science/hal-02564721

Submitted on 5 May 2020

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

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

Désidentification de comptes-rendus hospitaliers dans une base de données OMOP

N. Paris[1,2,3], M. Doutreligne[1,4], A. Parrot[1], X. Tannier[3]

[1]*WIND-DSI, AP-HP, Paris, France*
[2]*LIMSI, CNRS, Orsay, France*
[3]*LIMICS, INSERM, Paris, France*
[4]*Direction de la recherche, des études, de l'évaluation et des statistiques*

Résumé

En médecine, la recherche sur les données de patients vise à améliorer les soins. Pour préserver la vie privée des patients, ces données sont usuellement désidentifiées.

Les documents textuels contiennent de nombreuses informations présentes uniquement dans ce matériel et représentent donc un attrait important pour la recherche. Cependant ils représentent aussi un challenge technique lié au processus de désidentification.

Ce travail propose une méthode hybride de désidentification évaluée sur un échantillon des textes de l'entrepôt de données de santé de l'Assistance Publique des Hôpitaux de Paris.

Les deux apports principaux sont des performances de désidentification supérieures à l'état de l'art en langue française, et l'implémentation d'une chaîne de traitement standardisée librement accessible implémentée sur OMOP-CDM, un modèle commun de représentation des données médicales largement utilisé dans le monde.

Keywords:

Traitement Automatique des Langues, Standards de Référence

Introduction

L'Assistance Publique des Hôpitaux de Paris (APHP) met en place un entrepôt de données de Santé (EDS) qui colligera à terme l'ensemble des données de soin des patients. Outre l'aide au pilotage opérationnel, les cas d'usage prévoient un accès facilité aux données à des investigateurs de l'APHP dans le cadre de projets de recherche validés par un conseil scientifique et éthique (CSE) interne. La Commission Nationale de l'Informatique et des Libertés (CNIL) s'est prononcée favorablement pour la création de l'EDS de l'APHP. Des garanties sécuritaires légitimes ont été mises en place, en particulier la nécessité de la désidentification des données (ie : retirer les informations personnelles directement identifiantes). Si la désidentification des données médicales structurées est un processus maitrisé, celle des documents textuels est un challenge technique.

Les documents textuels hospitaliers (DTH) - comptes-rendus, lettres libres, synthèses de réunion - contiennent des informations cruciales et non présentes de manière structurées dans les dossiers patients informatisés (DPI). Si les DTH contiennent l'essentiel des informations médicales produites au cours d'une hospitalisation, leur exploitation pour la recherche médicale nécessite l'utilisation de traitement automatique des langues (TAL). Or, les DTH contiennent aussi des informations personnelles qui doivent être retirées pour permettre l'accès à ce contenu par les chercheurs en TAL en garantissant la protection de la vie privée des patients.

Les Etats-Unis ont été les premiers à définir des entités directement identifiantes (EDI) devant être retirées des DTH dans le cadre de l'Health Insurance Portability and Accountability Act (HIPAA). Dans le cadre de notre projet et parmi les 18 variables définies par le HIPAA, la CNIL s'est prononcée sur la désidentification de 8 EDI seulement. Ce nombre réduit de variables s'explique par le contexte de la recherche multi-centrique à l'APHP où les données restent dans un environnement sécurisé ainsi que par le fait que les traitements réalisés soient sous la responsabilité d'un investigateur salarié de l'APHP.

Des initiatives analogues ont été menées en langue française. L'outil Medina [1], accessible gratuitement, est un outil à base de règles. L'outil ALADIN [2] (code source propriétaire) est aussi à base de règles. L'outil en langue anglaise NeuroNER [3] est à base de réseaux de neurones. Les performances comparées de ces outils indiquent que les méthodes à base de réseaux de neurones surpassent les méthodes à bases de règles et pourraient être testées sur la langue française.

Le modèle commun de bases de données de recherche médicale Observational Medical Outcomes Partnership (OMOP)[1] remporte une adoption croissante dans le milieu médical. Ainsi OMOP collige 682 millions d'enregistrements patient du monde entier [4] et rend possible des analyses standardisées internationales grâce à la standardisation syntaxique (langage d'analyse et structuration des données) et sémantique (terminologies) des données médicales hospitalières. Les travaux du groupe de travail OMOP-NLP[2] dirigés par Hua Xu ont doté le modèle OMOP d'une table dédiée au TAL (nommée NOTE_NLP). Parallèlement l'EDS de l'APHP a progressivement normalisé l'ensemble de ces données (structurées ou non) dans ce format. Nous avons donc tenté de tirer partie des dernières avancées de la communauté OMOP dans le cadre de la désidentification des DTH de l'APHP.

Le premier apport de ce travail est d'avoir adapté les travaux de l'état de l'art en langue anglaise à la langue française. Le second apport de ce travail est d'avoir bâti une chaîne de traitement sur un modèle commun et international de représentation des données médicales. Le dernier apport de ce travail est la mise à disposition à la communauté médicale de l'ensemble des programmes généré ssur gestionnaire de versions de code source en ligne[3] (hormis les données

1 https://github.com/OHDSI/CommonDataModel/wiki
2 https://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:nlp-wg
3 https://framagit.org/interchu

d'apprentissage et le modèle neuronal entrainé pour cause de protection de la vie privée des patients).

Méthodes

Campagne d'annotation

NeuroNER avait utilisé un jeu d'apprentissage annoté de 2000 DTH. Dans ce travail, 3342 DTH ont été tirées au sort à partir de l'ensemble des DTH de l'APHP de manière à respecter les proportions des classes présentes (Table 1.), avec un nombre minimal de 15 documents pour les classes peu représentées (42 classes au total). Les DTH des patients étant opposés à la recherche (un formulaire en ligne permet aux patients de l'APHP de s'opposer à la recherche médicale sur données[4]), les personnalités publiques ainsi que les personnels hospitaliers APHP ont étés écartés.

Les DTH ont été pré-annotés avec le système à base de règles et de connaissances (cf. Infra) afin de simplifier la tâche des annotateurs. Les DTH ont étés séparés en 43 paquets destinés aux annotateurs. Dans chacun des paquets, 8 DTH étaient communs pour permettre la mesure de l'accord inter-annotateur. Ces données ont été placées sur un serveur intranet sécurisé sur lequel l'outil d'annotation Brat[5] a été installé.

Les annotateurs ont été sélectionnés parmi les membres de l'équipe de la direction des systèmes d'information de l'APHP et ont signé une clause de confidentialité. 15 personnes se sont portées volontaires (informaticiens, chefs de projets, internes en santé publique). Une formation de 30 minutes sur un jeu de démonstration a été réalisée afin de mettre en pratique le guide d'annotation et de prendre en main l'outil d'annotation Brat. L'annotation s'est déroulée dans une salle informatique et les échanges ont pu être menés pour partager les problématique et échanger sur les bonnes pratiques. L'ensemble des DTH a été annoté sur deux matinées de 4 heures avec le guide d'annotation ci-dessous:

L'étude d'accord inter-annotateur a permis d'écarter 2 annotateurs soit 160 documents portant le total de DTH finalement exploités à 3182. Un jeu de test de 286 documents respectant la distribution des types de DTH a été selectionné parmi ces derniers. Il a été reparcouru et réannoté minutieusement par un expert pour perfectionner le jeu de test.

Guide d'annotation : EDI

- ***id***: identifiant numérique national, de patient, de séjour
- **date**: référence explicite (mois, jour, année) hors durée
- **mail**: Adresse électronique
- **tel**: téléphone, fax
- **nom**: noms, prénoms, initiales (patient, professionel de santé, famille...)
- **adresse**: adresse postale, designation d'un lieu
- **ville**: nom de ville
- **zip**: code postal

Base de règles et connaissances

Un outil à base de règles et de connaissance a été créé. La réutilisation de l'outil Medina a été envisagée, puis écartée car les sorties de l'outil sont des textes balisés en XML. Il a été choisi de ne pas modifier les textes, mais de produire un fichier associé d'annotation Brat pour garantir la reproductibilité et simplifier la maintenance.

Règles

La version française du tokeniseur Stanford-parser[6] a été utilisée pour la segmentation en phrases. Le modèle entrainé sur le French Treebank corpus par Nicolas Hernandez a été utilisé [5]. Cette étape est utile pour l'outil Heideltime [6] qui découvre les dates et propose une normalisation. Heideltime a été enrichi de règles spécifiques aux données hospitalières françaises. Par ailleurs, des expressions régulières capturantes avec déclencheurs ont été produites pour récupérer les éléments semi-structurés les plus courants (téléphones, dates, identifiants). La recherche des informations issues de la base de données s'appuie sur une recherche par distance de Levenstein pondérée et normalisée par le clavier Azerty pour tolérer une certaine distance avec l'orthographe réel. L'UIMA-dict[7] a été amélioré pour normaliser les accents et appliquer une lemmatisation via le snowball français. Les briques choisies étant toutes écrites en java, la chaine de traitement a été implémentée via Apache UIMA/DKpro[8] et les calculs sont distribués sur une grappe de calcul via Apache Spark.

Connaissances

Les tables OBSERVATION / LOCATION / LOCATION_HISTORY fournies par OMOP contient les EDI structurés des patients de manière détaillée et historisée. Ils sont extraits et transmises au programme pour accompagner chaque DTH. OMOP a selectionné SNOMED comme terminologie de référence pour les données démographiques. Un dictionnaire de villes a été constitué à partir de la liste des villes françaises les plus peuplées ainsi que les villes d'Ile-De-France. Les villes qui sont des noms communs (ex: Plaisir, Le Port) ont été retirées par revue manuelle pour limiter les faux positifs sur cette classe.

Réseau de neurone

L'architecture mise en oeuvre est celle décrite par Lample et al. [7] avec réseau de neurone récurrent (bi-LSTM) associé à un CRF (Conditional Random Field). Un plongement lexical de mots à 300 dimensions a été produit sur 4 millions de documents textuels médicaux de l'APHP avec Word2Vec [8] puis réduit par PCA (Principal Component Analysis) à 50 dimensions. Le calcul des embeddings est un processus couteux en temps de calcul. La réduction de dimension par PCA permet de faire varier la dimension des embeddings en ne les calculant qu'une seule fois. Nous avons également fait des entraînements avec certaines petites tailles d'embedding sans observer de grands écarts de performances avec des embeddings 300 réduits par PCA. Le jeu annoté a été séparé en un jeu d'entrainement de 2589 DTH (1017098 mots), un jeu de développement de 307 DTH (125258 mots) et un jeu de test de 286 DTH (114415 mots), tous conformes aux distributions des types de DTH de l'EDS.

Gestions des résultats dans OMOP

Les comptes rendus sont enregistrés dans la table NOTE d'OMOP sous forme de chaine de caractères (VARCHAR), et

4 http://recherche.aphp.fr/eds/droit-opposition/
5 https://brat.nlplab.org/
6 https://nlp.stanford.edu/software/lex-parser.shtml
7 https://uima.apache.org/downloads/sandbox/DictionaryAnnotatorUserGuide/DictionaryAnnotatorUserGuide.html
8 https://dkpro.github.io/dkpro-core/

l'ensemble des caractères spéciaux sont remplacés par des espaces. Chaque CR est typé via une terminologie locale, et aligné si possible sur le code standard international LOINC. Les CR sont aussi liés aux patients ainsi qu'aux séjours hospitaliers.

Les patients ont de nombreuses données structurées médicales, ainsi que des données démographiques et identifiantes. Ces dernières sont intégrées dans la table OBSERVATION. A nouveau, chaque EDI est typé avec la terminologie locale utilisé à l'APHP et via la terminologie standard définie par OMOP : LOINC, SNOMED ou RxNORM selon les cas. La table NOTE_NLP d'OMOP recueille les résultats des deux méthodes à raison d'une ligne par EDI repéré. En plus des empans des EDI recueillis (index de début et de fin de l'entité nommée), cette table contient des métadonnées comme les dates d'extraction (date), le type d'algorithme utilisé (nlp_system), la date normalisée (temporal_term). Les concepts extraits sont reliés aux terminologies standards SNOMED et locales.

Dans ce contexte, le coût de la transformation d'un jeu de données au format OMOP est peu élevé puisque le modèle relationnel est simple et adapté aux données médicales. L'effort principal concerne l'alignement sémantique aux terminologies standards. Ici, l'alignement terminologique se résume à caractériser les EDI (table OBSERVATION) ainsi que les classes de documents (facultatif).

Chaîne de remplacement

La CNIL a suggéré une approche de remplacement des EDI plutôt qu'une approche de retrait. Il s'agit de remplacer de manière cohérente les EDI par des candidats issus de la base de données auxquels sont ajoutés des valeurs fictives. L'objectif est qu'un faux négatif laissé par la chaine de traitement soit ambigu pour renforcer les effets du traitement. Les dates sont décalées de manière homogènes pour un patient donné. Cela conserve donc les délais nécessaires à certaines études. Enfin, la CNIL a suggéré que les traitements de remplacements puissent être adaptés aux besoins et au contexte (+/- de remplacements).

Un programme dédié aux remplacements s'appuie sur les résultats présents dans la table NOTE_NLP issus des algorithmes UIMA et neuronaux. En fonction du contexte, l'ensemble ou partie des EDI peut être prise en compte. En cas de désaccord sur certains EDI entre les deux méthodes, le programme privilégie la base de connaissance. Les dictionnaires de candidats sont issus de la table OBSERVATION et augmentée de candidats fictifs générés par l'outil Python Faker[9] ou de dictionnaires publiques. Les EDI de remplacement sont tirés aléatoirement mais de manière reproductible pour un patient donné. Si dans un DTH précédent, le patient Jean avait été renommé en Paul, il le sera dans tous ses DTH et cela pour toutes les classes de EDI.

Résultats

Annotations

Table 1 – Principales classes de documents et leur nombre

Nombre	*Libellé Local*	*Code LOINC*	*Libellé LOINC*
608	Lettre consultation	-	-
366	CR consultation	11488-4	CR ou fiche de consultation ou de visite
274	CRH Service	11493-4	CR hospitalier (séjour)
222	Document Importé	-	-
189	Lettre Divers	-	-
150	CR de Jour	-	-
100	CR Opératoire	34874-8	CR Opératoire
91	Lettre CR	-	-
62	Bilan médical Initial	46241-6	CR d'admission

La table 1 représente un échantillon des 42 classes de document avec leur libellés tels que présents dans l'entrepôt associé si possible avec la terminologie standard définie par OMOP dans le cas des DTH: LOINC.

Table 2 – Accord inter-annotateurs par type

EDI	F1
date	.967
id	.991
nom	.944
ville	.740
Total	.963

L'accord inter-annotateur entre chaque pair d'annotateur a une f-mesure moyenne de 0.963 ce qui indique une bonne qualité d'annotation. L'accord inter-annotateur entre les différents EDI est donc globalement très bon hormis pour les villes. Les annotateurs ont été embarassés avec les hôpitaux incorporant les noms des villes, les noms d'hôpitaux n'étant pas à annoter contrairement aux noms de villes. En outre, le nombre de villes présentes dans les 8 documents selectionnés pour mesurer l'accord inter-annotateur était faible.

Les figure 1. et 2. représentent l'accord inter-annotateurs pour chacun des 43 paquets totaux. Deux paquets d'annotations (13, 21) ont un taux plus faible et ont été exclus pour améliorer la qualité globale du jeu annoté. On remarque que la seconde session a un accord inter-annotateur nettement supérieur à la première (Figure 2).

9 https://faker.readthedocs.io/en/latest/providers/faker.providers.python.html

Figure 1– Matrice d'accord inter-annotateurs (session 1)

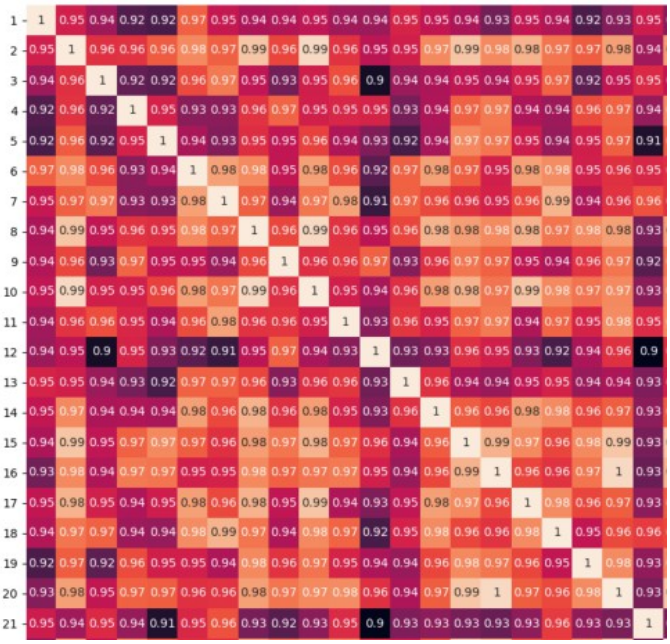
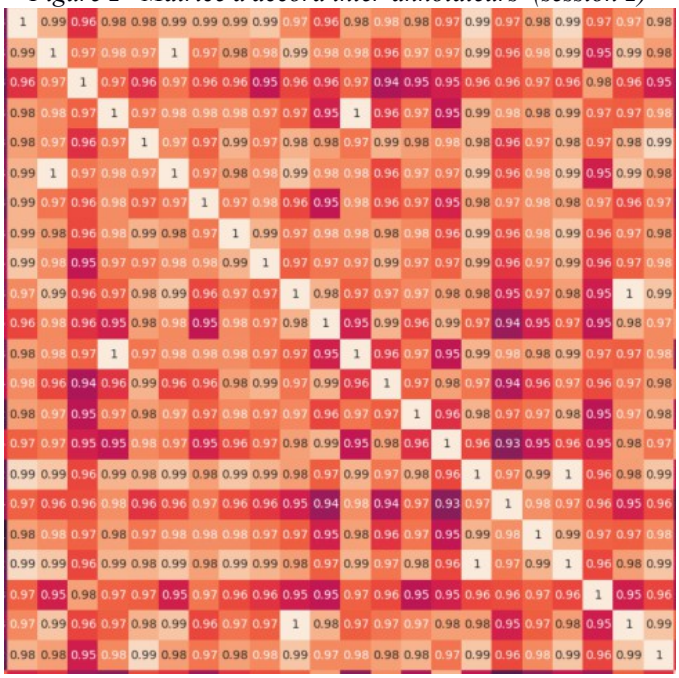


Figure 2– Matrice d'accord inter-annotateurs (session 2)



Performances

Table 3 – Performances comparative réseau de neurone (n) et hybride (h)

Type	N	Precision		Rappel		F1	
		n	*h*	*n*	*h*	*n*	*h*
adresse	85	.940	**.974**	**.951**	.927	.946	**.950**
date	2340	**.981**	.937	.941	**.992**	.961	**.964**
mail	4	.750	1	.750	**1**	.750	**1**
id	96	.978	**.989**	.979	**1**	.978	**.994**
nom	1884	**.983**	.971	.932	**.976**	,957	**.974**
tel	48	**.925**	.869	,948	**.952**	**,936**	.909
ville	168	.948	**.954**	,817	**.913**	,878	**.933**
zip	81	.987	.987	1	1	.993	.993
Total	4706	**.979**	.953	.935	**.981**	.957	**.967**

La méthode hybride surpasse la méthode neuronale en terme de rappel mais lui est infèrieure en terme de précision. Globalement l'hybridation augmente d'un point la F1.

Table 3– Performances comparees avec Medina (m), Aladin (a), NeuroNER (n) et hybride (h)

Type	F1			
	m	*a*	*n*	*h*
adresse	.182	.497	.898	**.950**
date	.910	.976	**.993**	.964
mail	-	.976	-	**1**
id	-	-	.988	**.994**
nom	.900	.827	**.992**	.974
tel	**1**	.895	.993	.909
ville	.579	-	-	**.933**
zip	**1**	-	**1**	.993

Le type adresse se démarque dans le cas de la méthode hybride. On remarque globalement un net avantage aux méthodes a base de reseau de neurones.

Discussion

Annotations

Les DTH sont très nuancés avec 42 types. Cependant deux tiers sont représentés par environ 10 classes seulement (présentés en Table 1). Le taux d'alignement sémantique à la terminologie de référence LOINC est faible et reflète une flexibilité des outils de recueil pour répondre aux besoins spécifiques des différents services hospitaliers.

Le taux d'accord inter-annotateur reflète une bonne qualité générale des annotations. Cependant le fait que 2 annotateurs sur les 15 sortent du lot laisse supposer des ambiguïtés dans le guide d'annotation. Le fait que la seconde séance ait un meilleur accord inter-annotateur peut s'expliquer par des phases d'échanges entre les deux séances. Deux types d'annotation ont posé des problèmes. En premier lieu les dates, qui ont pu parfois être confondues avec les durées, les saisons ou les dates de décrets qu'il fallait écarter. En second lieu, les villes ont le plus faible score inter-annotateur. Les 39 hôpitaux de l'APHP sont souvent liés à des noms de ville ou des noms (Hôpital Saint Antoine...) et cela a visiblement entrainé une confusion parmi les annotateurs. L'analyse des faux positif du réseau de neurone montre que cela introduit de la confusion. Présenter les hôpitaux aurait pu permettre d'améliorer les résultats.

Modèle hybride

L'inférence des EDI est 1000 fois plus rapide via le réseau de neurone que via une chaine UIMA. Cependant ces performances sont mitigées par la capacité à distribuer la chaine UIMA sur une grappe de serveur via Apache Spark. L'utilisation du modèle neuronal seul est possible cependant, le moteur de connaissance/règle apporte plusieurs avantages. En premier lieu, il augmente le rappel ce qui est le principal objectif d'une chaine de désidentification. En second lieu, il est plus flexible lorsque le contexte évolue et permet d'enrichir les bases de connaissances au fil du temps là où le réseau de neurones nécessite une nouvelle phase d'annotation et d'apprentissage. Une approche évolutive pour le réseau de neurone consiste à effectuer des petites annotations chaque année et faire évoluer la base d'entrainement. Enfin, le moteur de règle apporte la normalisation des dates qui offre la possibilité d'appliquer un décalage temporel lors de la phase de transformation des textes.

Comparaison avec l'état de l'art

Le contexte de ce travail est plus complexes que les chaines de désidentifications comparées. L'hétérogénéité des documents est bien plus importante. Les systèmes comparés traitent entre 1 et 4 services différents alors que le présent travail traite 800 services repartis sur 39 hôpitaux. De plus, les systèmes comparés traitent une seule classe de document alors que le présent travail traite 42 classes de documents. Dans le cas de NeuroNER, il s 'agit de la langue anglaise qui est plus dotée en terme d'outils de TAL. Si les EDI recherchés ne sont pas entièrement les mêmes, les types comparables montrent que le présent travail surpasse les performances des outils existants sur le français. La méthode hybride offre des performances proches de celles de NeuroNER, et le surpasse sur la classe des adresses.

Utilisation d'OMOP

La transformation au format OMOP est nécessaire pour pouvoir utiliser cette chaine de traitement - du point de vue syntaxique avec les tables OBSERVATION, NOTE et NOTE_NLP et point de vue sémantique avec l'utilisation de LOINC et de la SNOMED. Moyennant ce coût de transformation, notre travail présente plusieurs avantages. Les différentes étapes de la chaine de traitement ainsi que les résultats intermédiaires étant stockés dans une base de données, cela permet donc de réaliser des statistiques, des extractions, d'historiser, et de surveiller le bon fonctionnement au jour le jour dans un contexte de production de recherche. En outre, les données produites sont en lien avec des concepts standardisés comme définis par OMOP, mais aussi avec le reste du dossier patient, ce qui permet l'implémentation de règles de qualité fines et partageables d'une instance OMOP a l'autre, au niveau national ou international. Enfin, la chaine est compatible avec des fréquences de rafraichissement itératifs permettant de répondre aux enjeux de la médecine moderne et donc d'améliorer les soins au quotidien sur la base de données fraiches. En effet les entrepôts de données de santé enrichissent, consolident et normalisent l'information médicale via des algorithmes sur la base de laquelle d'autres algorithmes sont élaborés et validés. C'est donc dans les mêmes modalités que ces algorithmes doivent faire l'inférence; et il est donc nécessaire que les entrepôts soient raffraichis au plus vite. Par ailleurs, différents connecteurs aux base de données JDBC peuvent être utilises en entrée et sortie et de simples fichiers csv pourraient être utilisés, sans nécessiter une instance OMOP pleinement fonctionnelle. Ceci n'exclut donc pas les centres n'ayant pas fait le travail de standardisation de leurs bases de données au format OMOP.

Interopérabilité et partage d'algorithmes libres

Ce travail initie le partage d'algorithmes standards puisque construit sur le modèle standard OMOP de représentation des données en médecine. Nous espérons que ce travail démontera l'intérêt de standardiser les données des entrepôts de données de santé français et internationaux dans un modèle commun pour ensuite pouvoir bénéficier d'une multitude d'application, de processus de qualité et d'algorithmes standards et libres.

Nous espérons aussi participer à l'amélioration de l'interopérabilité dans le domaine de la santé via l'utilisation de modèle standard (OMOP) et d'API standard comme HL7-FHIR.

Conclusion

Les réseaux de neurones offrent des résultats supérieurs aux autres méthodes existantes moyennant un grand nombre de documents annotés. L'hybridation avec des méthodologies à base de règles et de connaissances permet d'augmenter le rappel, métrique importante pour cette tâche. La nature personnelle des donnée s d'apprentissage rend nécessaire la mise en place d'un jeu de donnée annoté, et rend impossible la diffusion publique du modèle appris.

L'implémentation de la chaîne de traitement sur le modèle OMOP, permet de produire des comptes-rendus désidentifiés directement exploitables dans une base de données de recherche. La capacité à rafraichir régulièrement la base de donnée offre la possibilité d'un retour des algorithmes au patient dans le cadre du soin. Enfin ce choix simplifie sa réutilisation, réévaluation et diffusion dans d'autres centres hospitaliers ayant un jeu de données au format standardisé.

Remerciements

Aucun financement ou bourse n'a financé ce travail.

Nous remercions Cyril Grouin, Aurélie Névéol, Franck Dermoncourt et Peter Szolovits pour les échanges et conseils sur cette thématique.
Nous remercions la communauté OHDSI pour leurs échanges réguliers.

References

[1] C. Grouin, A. Rosier, O. Dameron, P. Zweigenbaum. "Une procédure d'anonymisation a deux niveaux pour créer un corpus de comptes rendus hospitaliers." Risques, Technologies de l'Information pour les Pratiques Médicales. Informatique et Santé, vol 17, pp. 23-24 (2009)

[2] Q. Gicquel et al "Evaluation d'un outil d'aide à l'anonymisation des documents médicaux basé sur le traitement automatique du language naturel." Systèmes d'information pour l'amélioration de la qualité en santé. Informatique et Santé, vol 1., pp. 165-176 (2011)

[3] F. Dernoncourt, & J. Young Lee & O. Uzuner & P. Szolovits. "De-identification of Patient Notes with Recurrent Neural Networks." Journal of the American Medical Informatics Association : JAMIkA (2016)

[4] G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Re- ich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. Wong, P. R. Rijnbeek, J. van der Lei, N. Pratt, G. N. Noren, Y. C. Li, P. E. Stang, D. Madi- gan, P. B. Ryan, "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers." Stud Health Technol Inform, vol. 216, pp. 574–578 (2015).

[5] F . Boudin, N. Hernandez."Détection et correction automatique d'erreurs d'annotation morpho-syntaxique du French TreeBank." Actes de la conférence conjointe JEP-TALN-RECITAL 2012 - Traitement Automatique des Langues Naturelles (TALN), France (2012)

[6] J. Strotgen, M. Gertz. "HeidelTime: high quality rule-based extraction and normalization of temporal expressions." Proceedings of the 5th International Workshop on Semantic Evaluation; (2010)

[7] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, & C. Dyer (n.d.). "Neural Architectures for Named Entity Recognition", pp. 260–270 (2016)

[8] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality", Advances in Neural Information Processing Systems 26, pp. 848-854 (2013)

Adresse pour la correspondance:

M. Nicolas Paris : nicolas.paris@aphp.fr