HAL open science

Modèles préservant la confidentialité des données par mimétisme pour la reconnaissance d'entités nommées en français

Nesrine Bannour, Perceval Wajsbürt, Bastien Rance, Xavier Tannier, Aurélie Névéol

To cite this version:

Nesrine Bannour, Perceval Wajsbürt, Bastien Rance, Xavier Tannier, Aurélie Névéol. Modèles préservant la confidentialité des données par mimétisme pour la reconnaissance d'entités nommées en français. Journée d'étude sur la robustesse des systemes de TAL, ATALA, Nov 2022, Paris, France. hal-04013420

HAL Id: hal-04013420

https://hal.science/hal-04013420

Submitted on 10 Mar 2023

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.

Modèles préservant la confidentialité des données par mimétisme pour la reconnaissance d'entités nommées en français

N. Bannour[1] P. Wajsbürt[2] B. Rance[3,4,5] X. Tannier[2] A. Névéol[1]

(1) Université Paris-Saclay, CNRS, LISN, 91405 Orsay cedex, France
(2) Sorbonne Université, Inserm, Université Sorbonne Paris Nord, LIMICS, 75006 Paris, France
(3) INSERM, CRC, UMRS 1138, Université de Paris, Université Sorbonne Paris Cité, 75006 Paris, France
(4) Assistance Publique - Hôpitaux de Paris, Hôpital Européen Georges Pompidou, 75015 Paris, France
(5) HeKA, Inria Paris, 75006 Paris, France

RÉSUMÉ

Contexte – Les Dossiers Électroniques Patient (DEPs) présentent un fort potentiel pour améliorer la recherche clinique. Cependant, la plupart des données contenues dans les DEPs sont en format texte brut (Fu *et al.*, 2020). De plus, jusqu'à 80 % des informations cliniques cruciales ne sont disponibles que sous forme de texte non structuré (Escudié *et al.*, 2017; Jouffroy *et al.*, 2021). Dans ce projet, nous abordons l'extraction d'information dans des compte-rendus cliniques en français, qui consiste à identifier des entités médicales tels que Maladie, Anatomie, Médicament, etc. Les modèles d'apprentissage profond offrent de bonnes performances pour cette tâche de Reconnaissance d'entités nommées (REN). Néanmoins, la disponibilité de données d'entraînement cliniques annotées est souvent limitée, en particulier pour les langues autres que l'anglais. En outre, le caractère confidentiel des textes cliniques limite la possibilité d'échange de données entre les institutions. En effet, le partage de données est difficile dans la pratique et est strictement encadré par des réglementations telles que le RGPD [1]. Ainsi, l'adaptation de modèles de REN appris sur des corpus privés à des corpus publics est nécessaire pour permettre le partage d'outils d'extraction d'information clinique. Ce résumé présente des travaux détaillés dans (Bannour *et al.*, 2022).

Objectifs – Dans cette étude, nous étudions l'apprentissage par mimétisme (*Mimic learning*) (Baza *et al.*, 2020) pour la REN dans les rapports cliniques écrits en français, en utilisant à la fois les jeux de données publics et privés. L'idée de l'apprentissage par mimétisme est d'annoter des données publiques non étiquetées à l'aide d'un *modèle enseignant* (*teacher model*) privé qui a été entraîné sur les données sensibles originales. Les données publiques nouvellement étiquetées sont ensuite utilisées pour entraîner des *modèles élèves* (*student models*). Ces *modèles élèves* peuvent être partagés sans révéler les données sensibles d'origine ou exposer le modèle privé directement construit avec ces données. Notre but est de proposer une architecture de modèles préservant la confidentialité des données qui permettent aux institutions hospitalières de générer des modèles partageables, lorsqu'aucun corpus annoté n'est disponible publiquement.

Méthodologie – La Figure 1 illustre l'approche que nous proposons. Les compte-rendus cliniques sensibles sont utilisés pour entraîner un *modèle enseignant*. Plusieurs études (Chang & Li, 2018; Boulemtafes *et al.*, 2020) ont indiqué qu'il est possible de reconstruire approximativement une partie des données d'entraînement en observant simplement les prédictions. Par conséquent, le *modèle enseignant* privé ne sera utilisé que pour produire des annotations *Silver Standard* pour les données publiques, qui seront utilisées pour entraîner les *modèles élèves* partageables. En effet, le *modèle enseignant* restera privé et, comme pour les données sensibles, il ne pourra pas être partagé. Pour générer les *modèles élèves*, on utilise le *modèle enseignant* pour annoter le corpus public non étiqueté.

1. https://gdpr-info.eu/

De cette manière, nous créons un nouveau corpus annoté. Ce dernier est utilisé pour entraîner le *modèle élève*. Ce *modèle élève* partageable a pour objectif d'améliorer le transfert de connaissances sans révéler les informations de santé personnelles des patients. A partir d'un *modèle enseignant* entraîné sur le corpus privé MERLOT (Campillos *et al.*, 2018), nous générons trois *modèles élèves* préservant la confidentialité entraînés sur trois corpus publics : DEFT (Cardon *et al.*, 2020), CAS (Grabar *et al.*, 2018) et CépiDC [2]. Pour entraîner ces modèles, nous augmentons les annotations *Gold Standard* dont nous disposons avec des annotations *Silver* générées par le *modèle enseignant*. Nous comparons nos modèles avec trois modèles de référence : le *modèle enseignant* privé, un modèle public entraîné sur un corpus clinique annoté disponible publiquement et un modèle utilisant des dictionnaires construits à partir des bases de connaissance UMLS et JeuxDeMots.

Résultats – Le tableau 1 présente une comparaison de nos *modèles élèves* préservant la confidentialité avec le modèle privé. Bien que les meilleurs résultats soient obtenus avec le *modèle enseignant* privé, avec un score F1 de 0,857, l'utilisation de ce modèle privé pour créer des annotations *Silver* sur le corpus public DEFT/CAS semble être une technique efficace pour améliorer les performances de la REN clinique sur des données publiques. En effet, les *modèles élèves* obtiennent de meilleures performances que les autres modèles publics. Selon le Groupe de travail européen sur la protection des personnes à l'égard du traitement des données à caractère personnel [3], les techniques de préservation de la confidentialité doivent être évaluées selon trois critères : (i) est-il possible d'identifier directement un individu (ii) est-il possible de relier diverses informations qui pourraient conduire à l'identification d'un individu et (iii) est-il possible de déduire des informations relatives à un individu. Nous évaluons ces risques et nous affirmons qu'aucune attaque potentielle ne pourrait révéler des informations sur des données privées sensibles en utilisant les annotations *Silver* générées par le *modèle enseignant* sur des données non sensibles publiquement disponibles. Par conséquent, notre solution offre un bon compromis entre la performance et la préservation de la confidentialité. Mesurer l'impact des expériences menées pourrait être la première étape vers une prise de conscience et un contrôle de leur impact environnemental. Le tableau 1 présente l'empreinte carbone en termes d'équivalent CO2 en grammes. L'émission de CO2 résultant de l'entraînement du *modèle enseignant* et de notre meilleur *modèle élève* CAS est estimée équivalente à 2,52 km parcourus en voiture.

	Précision	**Rappel**	**F-Mesure**	**Équivalent CO_2 (g.)**
Modèle privé (*MERLOT, enseignant*)	0.852	0.862	0.857	123
Modèle public (*DEFT*)	0.592	0.383	0.465	22
Dictionary-based Model (*JDM*)	0.153	0.062	0.089	-
Dictionary-based Model (*UMLS*)	0.246	0.168	0.200	-
Modèle élève (*DEFT*)	0.604	0.743	0.666	30
Modèle élève (*CAS*)	**0.628**	**0.806**	**0.706**	169
Modèle élève (*CépiDc*)	0.580	0.710	0.638	394

TABLE 1 – La performance de nos modèles sur le corpus de test

MOTS-CLÉS : Confidentialité, Dossier Électronique Patient, Mimic learning, Réseaux de neurones, Reconnaissance d'entités nommées.

2. http://www.cepidc.inserm.fr/
3. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

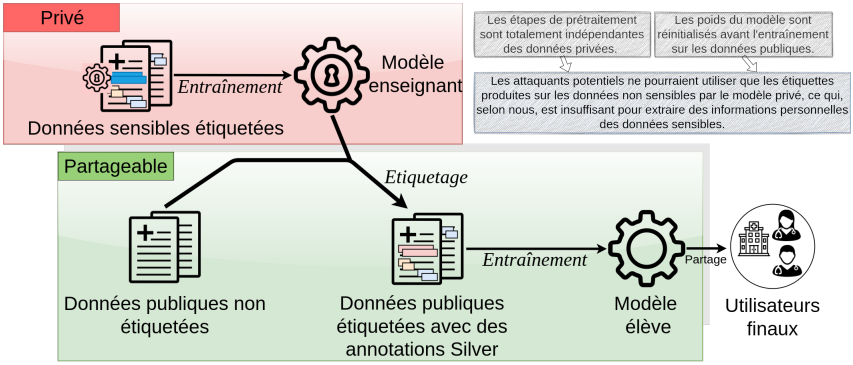

FIGURE 1 – Architecture des modèles préservant la confidentialité des données par mimétisme

---

Références

BANNOUR N., WAJSBÜRT P., RANCE B., TANNIER X. & NÉVÉOL A. (2022). Privacy-preserving mimic models for clinical named entity recognition in french. *Journal of Biomedical Informatics*, **130**, 104073. DOI : https ://doi.org/10.1016/j.jbi.2022.104073.

BAZA M., SALAZAR A., MAHMOUD M., ABDALLAH M. & AKKAYA K. (2020). On sharing models instead of data using mimic learning for smart health applications. In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, p. 231–236. DOI : 10.1109/ICIoT48696.2020.9089457.

BOULEMTAFES A., DERHAB A. & CHALLAL Y. (2020). A review of privacy-preserving techniques for deep learning. *Neurocomputing*, **384**, 21–45.

CAMPILLOS L., DELÉGER L., GROUIN C., HAMON T., LIGOZAT A.-L. & NÉVÉOL A. (2018). A french clinical corpus with comprehensive semantic annotations : development of the medical entity and relation limsi annotated text corpus (merlot). *Language Resources and Evaluation*, **52**(2), 571–601.

CARDON R., GRABAR N., GROUIN C. & HAMON T. (2020). Présentation de la campagne d'évaluation deft 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques. In *Actes de l'atelier Défi Fouille de Textes@JEP-TALN 2020 similarité sémantique et extraction d'information fine. Atelier DÉfi Fouille de Textes*, p. 1–13, Nancy, France : Association pour le Traitement Automatique des Langues.

CHANG S. & LI C. (2018). Privacy in neural network learning : Threats and countermeasures. *IEEE Network*, **32**, 61–67.

ESCUDIÉ J.-B., RANCE B., MALAMUT G., KHATER S., BURGUN A., CELLIER C. & JANNOT A.-S. (2017). A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease : a case study on autoimmune comorbidities in patients with celiac disease. *BMC medical informatics and decision making*, **17**(1), 1–10.

FU S., CHEN D., HE H., LIU S., MOON S., PETERSON K. J., SHEN F., WANG L., WANG Y., WEN A., ZHAO Y., SOHN S. & LIU H. (2020). Clinical concept extraction : A methodology review. *Journal of Biomedical Informatics*, **109**, 103526. DOI : https ://doi.org/10.1016/j.jbi.2020.103526.

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). CAS : French corpus with clinical cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, p. 122–128, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5614.

JOUFFROY J., FELDMAN S. F., LERNER I., RANCE B., BURGUN A. & NEURAZ A. (2021). Hybrid deep learning for medication-related information extraction from clinical texts in french : Medext algorithm development study. *JMIR Medical Informatics*, **9**.