Traitement de la langue naturelle pour une réponse rapide aux maladies émergentes: COVID-19

Neuraz Antoine[1] Lerner Ivan[1] Digan William[2]
Garcelon Nicolas[3] Tsopra Rosy[2] Rogier Alice[2] Baudoin David[2]
Burgun Anita[1, 2] Rance Bastien[2]

(1) INSERM, Centre de Recherche des Cordeliers, UMRS 1138, Université de Paris, Hôpital Necker Enfant Malade, Assistance Publique - Hôpitaux de Paris

(2) INSERM, Centre de Recherche des Cordeliers, UMRS 1138, Université de Paris, Hôpital Européen Georges Pompidou, Assistance Publique - Hôpitaux de Paris

(3) INSERM, Centre de Recherche des Cordeliers, UMRS 1138, Institut Imagine, INSERM U1163, Université Paris Descartes, Université de Paris, Paris, France

antoine.neuraz@aphp.fr

RÉSUMÉ

La fouille de données biomédicales dans les dossiers patients informatisés (DPI) a souvent été proposée comme méthode pour convertir les données non structurées vers les données structurées nécessaires pour la santé publique. Bien que cela ait souvent été suggéré (Elkin *et al.*, 2012), l'occasion ne s'était encore jamais présentée de pouvoir tester cette hypothèse en temps réel. Ainsi, la crise du coronavirus, malgré toutes ses tragédies, présente également l'opportunité d'améliorer l'informatique de santé publique. Durant la crise, l'APHP a mis en place une base de données au format OMOP CDM (Hripcsak *et al.*, 2015) contenant les données des DPI de tous les patients testés COVID-19. Voici un résumé de la méthode que nous avons utilisé pour traiter les textes clinique (Figure 1) : (1) un pré-traitement classique (*i.e.*, nettoyage du texte, détection des phrases) a été appliqué sur l'ensemble du dataset ; (2) l'extraction des noms de médicaments et des détails de prescription (dose, voie d'administration, fréquence, durée) a été effectuée à l'aire de modèles de deep-learning basés sur des embeddings contextuels de type BERT (Devlin *et al.*, 2018) fine-tuné sur 10M de textes cliniques et un modèle BiLSTM-CRF (Lample *et al.*, 2016) ($NLP_{medication}$) ; (3) l'extraction de phénotypes spécifiques associés au COVID-19 (*e.g.*, obésité, fumeur), de scores (*e.g.*, IGS2), et de mesures physiologiques (*e.g.*, Body Mass Index), a été effectuée via une liste d'expressions régulières spécialement développées ; (4) l'extraction de tous signes, symptômes, comorbidités présentes dans le Unified Medical Language System (UMLS) (Lindberg *et al.*, 1993), a été effectuée avec l'algorithme quickUMLS (Okazaki & Tsujii, 2010)). L'utilisation d'outils TAL a permis d'augmenter, de manière importante, la quantité d'informations, concernant les médicaments et les phénotypes, disponibles pour l'analyse. Le nombre de points de données pour les médicaments a été multiplié par 7.2 et le nombre de phénotypes par 15.2. Parmi les 84,966 dossiers présents dans la base EDS-COVID, 53% des patients avaient des informations sur les médicaments dans les textes cliniques contre seulement 23% dans les champs structurés. Pour les phénotypes spécifiques avec des codes CIM10 existant, l'information était disponible uniquement dans le texte libre pour une majorité de patients : 7,133/8,526 (83%) pour le diabète et 2,138/2,871 (74%) pour l'obésité. Certains items étaient absents des données structurées mais ont pu être récupérés via le TAL, comme l'agueusie ou l'anosmie, 2,449 et 2,732 patients respectivement. $NLP_{medication}$ a montré une F1-mesure brute à 93.8% (91.6% après normalisation) pour l'extraction des noms de médicament sur l'ensemble des sections ; 96.7%

(96% après normalisation) sur les sections traitement à l'admission et traitement de sortie.

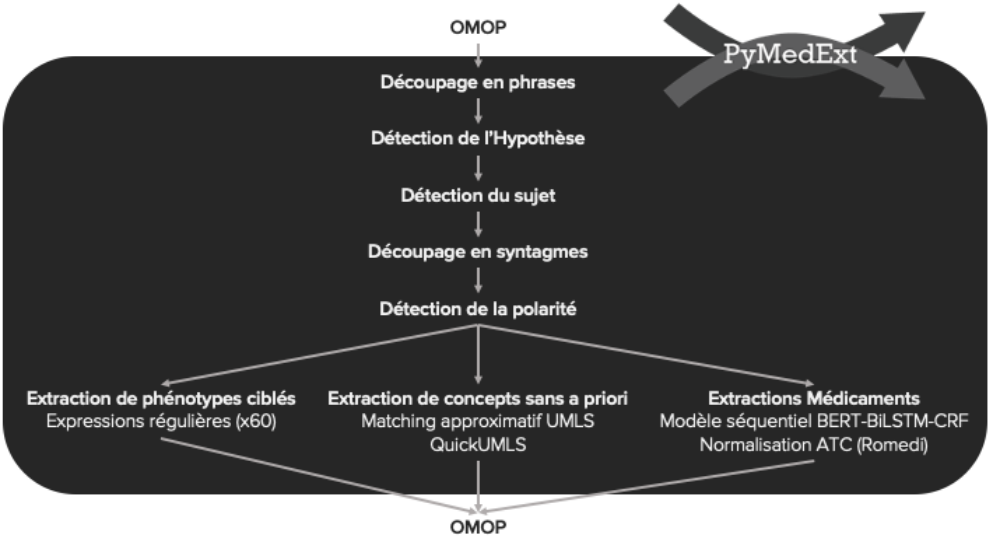

FIGURE 1 – Pipeline de traitement des textes cliniques

MOTS-CLÉS : Dossier patient informatisé, extraction d'information, COVID-19, deep-learning.

Références

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv :1810.04805 [cs]*.

ELKIN P. L., FROEHLING D. A., WAHNER-ROEDLER D. L., BROWN S. H. & BAILEY K. R. (2012). Comparison of natural language processing biosurveillance methods for identifying influenza from encounter notes. *Annals of Internal Medicine*, **156**(1_Part_1), 11–18.

HRIPCSAK G., DUKE J. D., SHAH N. H., REICH C. G., HUSER V., SCHUEMIE M. J., SUCHARD M. A., PARK R. W., WONG I. C. K., RIJNBEEK P. R., VAN DER LEI J., PRATT N., NORÉN G. N., LI Y.-C., STANG P. E., MADIGAN D. & RYAN P. B. (2015). Observational Health Data Sciences and Informatics (OHDSI) : Opportunities for Observational Researchers. *Studies in Health Technology and Informatics*, **216**, 574–578.

LAMPLE G., BALLESTEROS M., SUBRAMANIAN S., KAWAKAMI K. & DYER C. (2016). Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT 2016*, volume 5805, p. 260–270, San Diego, California, June 12-17, 2016 : ACL.

LINDBERG D. A., HUMPHREYS B. L. & MCCRAY A. T. (1993). The Unified Medical Language System. *Methods Archive*, **32**, 281–291.

OKAZAKI N. & TSUJII J. (2010). Simple and Efficient Algorithm for Approximate Dictionary Matching. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, p. 851–859, Beijing, China : Coling 2010 Organizing Committee.