PyMedExt, un couteau suisse pour le traitement des textes médicaux

William Digan[1, 2] Alice Rogier[1, 2] David Baudoin[1, 2] Bastien Rance[1, 2]
Antoine Neuraz[1, 3]
(1) INSERM, Centre de Recherche des Cordeliers, UMRS 1138. Paris, France.
(2) Hôpital Européen Georges Pompidou, AP-HP, 75015 Paris, France
(3) Hôpital Necker - Enfants malades, AP-HP, 75015 Paris, France
william.digan@institutimagine.org, alice.rogier@inria.fr,
david.baudoin@aphp.fr, bastien.rance@aphp.fr, antoine.neuraz@aphp.fr

RÉSUMÉ

Le paysage du traitement des textes est vertigineux pour les non-spécialistes du traitement des langues. La prise en main d'outils existants, la contribution à leur évolution et le partage de proposition locale sont rendus complexes par l'absence de formats d'annotateurs à la fois faciles d'utilisation et simplement déployables. Il existe cependant de nombreuses ressources d'intérêts développées par la communauté. Nous proposons PyMedExt, un couteau suisse pour l'annotation de textes cliniques. PyMedExt a pour objectif de fluidifier la mise en place et la communication entre les différents composants nécessaires en traitement des textes cliniques : les formats de représentation des annotations, les annotateurs et des pipelines simples.

PyMedExt est construit autour de trois contributions principales :

1. Formats d'annotation et convertisseurs. PyMedExt peut consommer en entrée des fichiers textes bruts, des flux FHIR de textes, mais aussi des fichiers dans des formats classiques de TAL (BioC, CoNLL...). Il peut prendre en charge les conversions de format de et vers ces formats, ainsi que vers une représentation interne exportable en JSON. Cette représentation interne étend le format BioC en permettant l'héritage entre entités. PyMedExt propose également l'envoi vers les outils d'annotation et de visualisation d'annotations BRAT (Stenetorp *et al.*, 2011) et Doccano (Nakayama *et al.*, 2018), ainsi que la possibilité d'exporter les textes et les annotations vers une base de données au format CDM OMOP (Hripcsak *et al.*, 2015).
2. Un format de représentation d'annotateurs. PyMedExt inclut un squelette d'annotateur qui peut être facilement étendu pour créer ses propres fonctionnalités, ou afin de développer un adaptateur (wrapper) pour des outils existants. Quelques exemples d'annotateurs PyMedExt natifs ou d'adaptateurs, dont QuickUMLS (Soldaini & Goharian, 2016) et HeidelTime (Strötgen & Gertz, 2010), sont disponibles sur le dépôt du projet (https://github.com/equipe22/pymedext_public). Le format d'annotateur est volontairement simple et peu contraignant pour favoriser une adoption par une communauté large.
3. Gestionnaire de pipelines simples. Enfin, PyMedExt propose un système de gestion de pipelines linéaires élémentaires d'annotateurs, permettant le déploiement pour un usage reproductible. PyMedExt est distribué sous la forme d'une bibliothèque Python, utilisable en ligne de commande pour les fonctionnalités de conversion, et embarquée dans un programme pour les fonctionnalités plus avancées (conversion, annotateurs et pipelines).

PyMedExt est utilisé par les équipes d'informatique de l'hôpital Necker et de l'HEGP pour standardiser les pratiques et simplifier le partage d'outils. Une bibliothèque d'annotateurs open-source publics au format PyMedExt est proposée en ligne https://github.com/equipe22/pymedext_core.

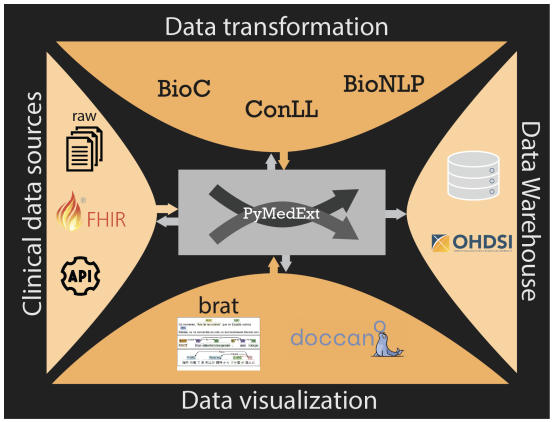

FIGURE 1 – Résumé graphique de PyMedExt

Références

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

HRIPCSAK G., DUKE J., SHAH N., REICH C., HUSER V., SCHUEMIE M., SUCHARD M., PARK R., WONG I., RIJNBEEK P., VAN DER LEI J., PRATT N., NORÉN G., LI Y., STANG P., MADIGAN D. & RYAN P. (2015). Observational Health Data Sciences and Informatics (OHDSI) : Opportunities for Observational Researchers. In *Stud Health Technol Inform*, volume 216, p. 574–8.

NAKAYAMA H., KUBO T., KAMURA J., TANIGUCHI Y. & LIANG X. (2018). Doccano : Text Annotation Tool for Human.

SOLDAINI L. & GOHARIAN N. (2016). QuickUMLS : a fast, unsupervised approach for medical concept extraction. In *SMedIR Workshop, SIGIR*.

STENETORP P., TOPIĆ G., PYYSALO S., OHTA T., KIM J.-D. & TSUJII J. (2011). BRAT. BioNLP Shared Task 2011 : Supporting Resources. In *Proceedings of BioNLP Shared Task 2011 Workshop*.

STRÖTGEN J. & GERTZ M. (2010). HeidelTime : High quality rule-based extraction and normalization of temporal expressions. In *SemEval '10 : Proceedings of the 5th International Workshop on Semantic Evaluation*, volume 216, p. 321–324.