Mieux comprendre le parcours de soin des patients à l'hôpital : extraction de connaissances temporelles dans des cas cliniques

Responsables

Anaïs Halftermeyer (anais.halftermeyer@univ-orleans.fr)	LIFO	Université d'Orléans
Silvia Federzoni (silvia.federzoni@univ-orleans.fr)	LLL	Université d'Orléans
Jean-Yves Antoine (jean-yves.antoine@univ-tours.fr)	LIFAT	Université de Tours
Anne-Lyse Minard-Forst (anne-lyse.minard@univ-orleans.fr)	LLL	Université d'Orléans

Résumé

Contexte scientifique

Le laboratoire d'Informatique de l'Université d'Orléans (LIFO) propose un stage en collaboration avec les laboratoires LLL de l'Université d'Orléans et LIFAT de l'Université de Tours, dans le cadre d'un financement de la fédération ICVL (Informatique Centre Val de Loire).

Le projet s'inscrit dans la continuité du projet APR-IA DOING et des recherches en cours localement sur l'extraction des connaissances (notamment des relations temporelles) dans le domaine médical et l'explicabilité des systèmes d'extraction des relations temporelles. Il vise à extraire des scénarios génériques des parcours de patients à partir d'un corpus de rapports de cas cliniques.

Lorsqu'une personne entre à l'hôpital, pour une durée parfois longue, son parcours (passages au sein des différents services, évolution de son état de santé, examens réalisés, etc.) est systématiquement documenté sous la forme de rapports textuels. Ces rapports décrivent la pathologie de la personne, ses symptômes, les différentes interventions réalisées ou encore les traitements administrés par les praticiens. Les rapports constituent ainsi une masse très riche d'informations anonymisées permettant par exemple de faire des analyses sur l'efficacité de certaines stratégies thérapeutiques, et de constituer une base d'apprentissage pour des applications de l'IA en santé. Pour ce faire, travailler directement sur des textes bruts se révèle insuffisant.

Afin de mieux cerner les parcours des patients, il est notamment important de pouvoir caractériser des relations temporelles au fil des rapports. Supposons, par exemple, que le rapport indique que ***à son arrivée*** la personne présentait **des symptômes (des douleurs intenses au mollet et une impotence fonctionnelle totale de ce dernier)**, qu'une **IRM** a été réalisée pour détecter la cause de **ces symptômes**, que **l'examen** a mis en évidence **un syndrome de loges** nécessitant **une intervention chirurgicale**, laquelle a conduit **à une immobilisation de la jambe *pendant 3 semaines***, ***puis*** à la disparition **des symptômes**. Afin d'exploiter ces informations pour tracer les parcours des patients, il serait utile d'annoter le texte pour préciser l'existence de relations (temporelles : avant/après, chevauchement d'événements, ect.) entre les termes en gras, mais également de préciser que les trois mentions du terme "symptôme" renvoient à la même chose (on parle alors en linguistique de coréférence entre ces trois mentions). La finalité de ce stage est précisément de permettre une annotation automatique de ces informations dans le texte.

Dans ce cadre, le ou la stagiaire appliquera et évaluera une approche exploratoire encore peu utilisée pour l'extraction de connaissances : l'analyse de séquences (Blanchard, 2019) à partir de données en langue naturelle. Les séquences sont "des listes ordonnées d'états ou d'événements, dont l'analyse permet d'identifier les régularités, les ressemblances et de construire des typologies de 'séquences-types'" (Robette, 2011). Pour ce qui est des cas cliniques, les séquences sont donc des représentations d'états ou d'événements cliniques (signes et symptômes, traitements, examens), auxquels sont associées des entités pouvant entretenir une relation de coréférence (relation qui existe entre deux expressions qui réfèrent à une même entité). Les séquences analysées comportent donc des dimensions complexes, telles les relations temporelles entre les événements et les phénomènes de coréférence, dont nous faisons l'hypothèse qu'une prise en compte simultanée favoriserait l'extraction des connaissances, à savoir l'identification de scénarios cliniques récurrents.

Le ou la stagiaire travaillera sur un corpus multilingue de cas cliniques (Magnini et al., 2021) annoté en entités cliniques et en relations temporelles. La première étape du travail commencera par une annotation automatique de la coréférence afin d'enrichir le corpus avec cette nouvelle couche d'annotation. À partir de l'ensemble des annotations en coréférence et en relations temporelles, le ou la stagiaire construira des séquences d'événements cliniques sur lesquelles il ou elle appliquera l'analyse des séquences. Le ou la stagiaire proposera ensuite des patrons formalisant les connaissances extraites et évaluera leur capacité à être interprétables, réutilisables et utiles pour la modélisation des parcours cliniques et des relations temporelles. Ce travail contribuera à la constitution d'une base de connaissances qui sera utile pour l'évaluation de systèmes d'extraction de relations temporelles. Le stage explorera notamment si et comment ces connaissances peuvent servir de leviers d'explicabilité des systèmes partiellement ou non explicables.

Travail à réaliser (estimation temporelle en semaines)

- **Phase 1 : découverte de la problématique et annotation automatique de la coréférence ($\mathbf{T0}$–$\mathbf{T0_{+3}}$)** —Un outil existant pour l'annotation automatique de la coréférence en français a déjà été testé sur un autre corpus de cas cliniques. Pour l'annotation automatique de la coréférence pour les autres langues, des systèmes multilingues existent mais n'ont pas été testés sur le corpus

des cas cliniques. Il s'agira de tester ces outils sur le corpus E3C[1] afin de produire une annotation automatique de la coréférence sur l'ensemble du corpus exploité. Cette étape permettra d'enrichir le corpus E3C avec une couche d'annotation supplémentaire générée automatiquement.

- **Phase 2 : mise en format des annotations, création des séquences et application de l'analyse des séquences ($T0_{+3}$–$T0_{+10}$)** —Dans un premier temps, il s'agira d'extraire les annotations existantes (entités cliniques, relations temporelles et coréférence) pour construire des séquences d'événements. Ensuite, il s'agira d'appliquer l'analyse des séquences. À partir des résultats obtenus, l'objectif est de formaliser les séquences représentatives des parcours des patients par des patrons exploitables comme unités de connaissance.

- **Phase 3 : évaluation des patrons extraits et profilage des parcours cliniques ($T0_{+10}$–$T0_{+18}$)** —Cette étape consistera à évaluer les patrons extraits et leur capacité à être interprétés et réutilisés comme unités de connaissance. L'évaluation sera principalement qualitative (et si possible quantitative) et portera sur la valeur explicative des patrons pour la modélisation des parcours cliniques. Cette étape pourra donner lieu à la constitution d'une base de connaissance qui pourrait servir de ressource à d'autres études portant sur les cas cliniques.

- **Phase 4 : étude du rôle de la connaissance extraite comme levier d'explicabilité pour les systèmes cibles ($T0_{+18}$–$T0_{+24}$)** —Cette étape consistera à étudier comment les connaissances extraites à partir des données peuvent contribuer à améliorer la compréhension des systèmes partiellement ou non explicables.

Résultats attendus

- Enrichissement du corpus E3C avec une couche d'annotation automatique de la coréférence ;
- Modélisation de scénarios génériques de parcours des patients ;
- Constitution d'une base de connaissances pour l'évaluation de systèmes partiellement ou non explicables.

Profil recherché

Ce stage s'adresse à un(e) étudiant(e) en master 1 ou 2 en informatique ou en master 2 en traitement automatique du langage, disposant de solides compétences en programmation Python. La maîtrise du langage R serait appréciée. Le ou la candidat(e) devra en outre posséder des connaissances linguistiques ou manifester un intérêt prononcé pour les problématiques liées au langage. Mais avant tout, on attend de la personne recrutée qu'elle présente un intérêt marqué pour la recherche, qu'elle ait une autonomie et un sens critique développés. Ce stage de découverte est donc proposé à des étudiants qui disposeraient d'un

[1]Le corpus E3C est disponible ici : `https://github.com/hltfbk/E3C-Corpus`.

excellent niveau académique, d'une curiosité scientifique affirmée et qui réfléchiraient à une orientation professionnelle future dans le domaine de la recherche.

Date et lieu de stage

La personne recrutée travaillera au sein du laboratoire LIFO (Université d'Orléans) où elle s'intégrera dans l'équipe Contraintes et Apprentissage (`http://www.univ-orleans.fr/lifo/equipes/CA/`). Elle travaillera en collaboration constante avec Anaïs Halftermeyer, de l'équipe Contraintes et Apprentissage du LIFO ainsi qu'avec Silvia Federzoni et Anne-Lyse Minard-Forst du laboratoire LLL (`https://www.univ-orleans.fr/fr/lll/le-laboratoire`) et Jean-Yves Antoine du laboratoire LIFAT (`https://lifat.univ-tours.fr/lifat-version-francaise/accueil`, Université de Tours).

Durée et période de stage - La durée du stage sera de 6 mois. Début de stage fin janvier / début février 2026 (à négocier avec la personne sélectionnée (avril 2026 au plus tard)).

Rémunération

La personne recrutée recevra une gratification mensuelle correspondant à la réglementation, à savoir 15% du plafond horaire de la sécurité sociale. À titre d'exemple, cette gratification représente un montant de 669,90 € pour un mois avec 22 jours ouvrés, et 609 € pour un mois avec seulement 20 jours ouvrés (jours fériés, par exemple). La personne recrutée participera aux réunions de l'équipe projet.

Contacts - Dépôts de candidature

Dépôt des candidatures par courrier électronique auprès de Silvia Federzoni (silvia.federzoni@univ-orleans.fr), avant le 15 décembre 2025, délai de rigueur. Merci de déposer :

- Un CV détaillé de vos activités passées ;
- Une lettre de motivation ;
- Vos relevés de notes des deux dernières années d'études.

Le cas échéant, un développement Python et/ou une lecture critique d'article scientifique pourront être demandés pour la sélection.

References

Blanchard, P. (2019). Sequence analysis. In *Encyclopedia of research methods*. Sage, London, p.a. atkinson, r.a. williams and a. cernat edition.

Magnini, B., Altuna, B., Lavelli, A., Speranza, M., and Zanoli, R. (2021). The E3C Project: Collection and Annotation of a Multilingual Corpus of Clinical Cases. In Monti, J., Tamburini, F., and Dell'Orletta, F., editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020 : Bologna, Italy, March 1-3, 2021*, Collana Dell'Associazione Italiana Di Linguistica Computazionale, pages 258–264. Accademia University Press, Torino.

Mathew, S., Peat, G., Parry, E., Sokhal, B. S., and Yu, D. (2024). Applying sequence analysis to uncover 'real-world' clinical pathways from routinely collected data: A systematic review. *Journal of Clinical Epidemiology*, 166.

Robette, N. (2011). *Explorer et décrire les parcours de vie: les typologies de trajectoires*. Les collections du CEPED. CEPED, Paris.