

ANNOTATION SÉMANTIQUE DE DOCUMENTS CLINIQUES PSYCHIATRIQUES FRANÇAIS FONDÉE SUR UNE ONTOLOGIE DE DOMAINE

Ons Aouina¹, Jacques Hilbey¹, Jean Charlet^{1,2}

5 juillet 2023

¹LIMICS (INSERM/Sorbonne Université /Sorbonne Paris Nord)

²AP-HP/DRCI

Introduction

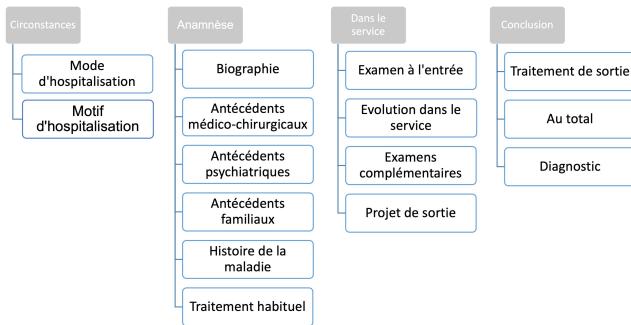
Annotation sémantique : Motivation

- PsyCARE¹ est un projet RHU qui vise à mettre en place une psychiatrie préventive et personnalisée dans le domaine de la schizophrénie
- L'annotation sémantique est le processus qui consiste à attacher à un document textuel ou à un autre contenu non structuré des métadonnées sur les concepts qui s'y rapportent (Neves et Ševa 2019)

1. <https://psy-care.fr/>

Contexte (1/2)

L'analyse des comptes rendus d'hospitalisation (CRH) permet d'obtenir des informations utiles sur la manière dont les événements médicaux affectent la progression de la psychose chez les patients



Défis à relever

Les défis sont liés à l'extraction d'informations lors du le traitement des comptes rendus cliniques psychiatriques français :

- Variabilité linguistique
- Manque de structure : les rapports cliniques non structurés n'ont pas de structure ou de format standardisé
- Absence d'informations ou informations incomplètes
- Absence de données annotées

L'objectif principal est de construire une chronologie complète de l'évolution de la maladie d'un patient à partir de son dossier médical.

Comme première brique de ce projet, nous avons construit :

- Un pipeline qui permet l'annotation sémantique des comptes rendus des patients à l'aide de l'outil **GATE** (Cunningham et al. 2014)
- Une évaluation manuelle du système d'annotation par deux annotateurs sur **50 CRHs**

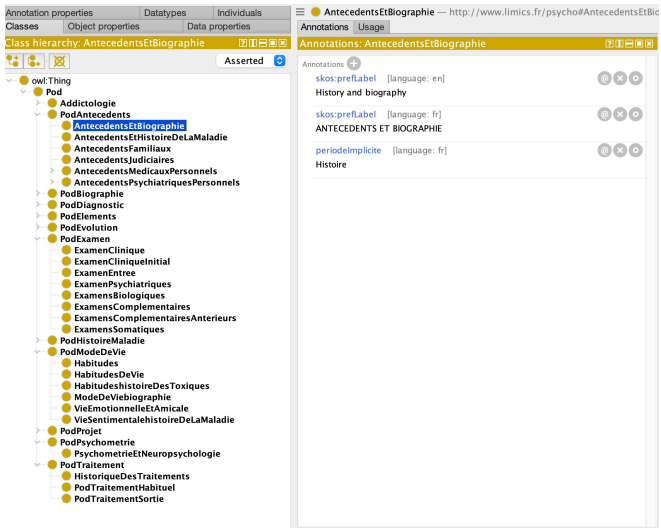
Matériels & Méthodes

Pour assurer l'annotation sémantique, nous avons utilisé les outils suivants :

- Les CRHs issus du projet français PsyCARE
- Ontologie de domaine (Hilbey, Aimé et Charlet 2022) développée dans le cadre de PsyCARE qui représente :
 - Aspects cliniques et psychiatriques, médicaments avec leurs codes ATC², imagerie, biologie, etc
 - Structure des sections des résumés de sortie des patients

2. Liste de classification médicale de l'OMS
<https://icd.who.int/browse10/2019/en>

Matériels



The screenshot displays a web interface for a class hierarchy and its annotations. The left pane shows a tree view under 'owl:Thing' with a sub-tree 'Pod'. The class 'AntecedentsEtBiographie' is highlighted. The right pane shows the annotations for this class.

Class hierarchy (Left Pane):

- owl:Thing
 - Pod
 - Addictologie
 - PodAntecedents
 - AntecedentsEtBiographie** (highlighted)
 - AntecedentsEtHistoireDeLaMaladie
 - AntecedentsFamiliaux
 - AntecedentsJudiciaires
 - AntecedentsMedicauxPersonnels
 - AntecedentsPsychiatriquesPersonnels
 - PodBiographie
 - PodDiagnostic
 - PodElements
 - PodEvolution
 - PodExamen
 - ExamenClinique
 - ExamenCliniqueInitial
 - ExamenEntree
 - ExamenPsychiatriques
 - ExamensBiologiques
 - ExamensComplementaires
 - ExamensComplementairesAnterieurs
 - ExamensSomatiques
 - PodHistoireMaladie
 - PodModeDeVie
 - Habitudes
 - HabitudesDeVie
 - HabitudeshistoireDesToxiques
 - ModeDeViebiographie
 - VieEmotionnelleEtAmicale
 - VieSentimentalehistoireDeLaMaladie
 - PodProjet
 - PodPsychometrie
 - PsychometrieEtNeuropsychologie
 - PodTraitement
 - HistoriqueDesTraitements
 - PodTraitementHabituel
 - PodTraitementSortie

Annotations (Right Pane):

Annotations: AntecedentsEtBiographie

- skos:prefLabel [language: en] History and biography
- skos:prefLabel [language: fr] ANTECEDENTS ET BIOGRAPHIE
- periodeImplicite [language: fr] Histoire

Figure 1 – Structure des sections des comptes rendus des patients

Méthodes : Processus d'annotation

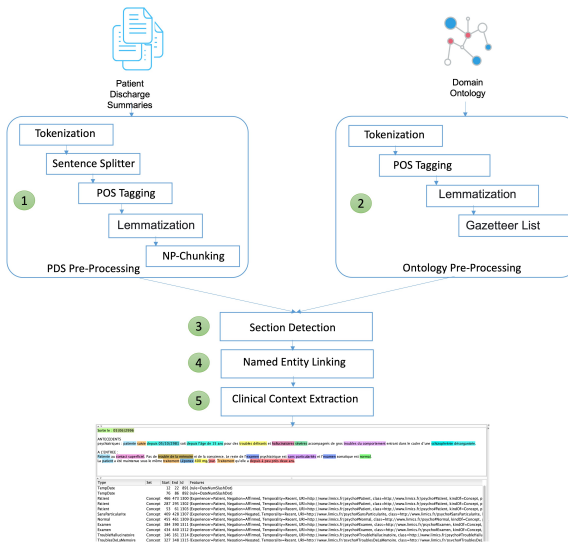


Figure 2 – Illustration du processus d'annotation sémantique fondé sur une ontologie

Résultats

- Un schéma d'annotation est créé pour évaluer les annotations du pipeline à l'aide de l'outil brat³
- Les entités extraites sont regroupées en 10 uniques **concepts d'ontologie de haut niveau** pour faciliter l'évaluation manuelle
- Deux annotateurs ont évalué les annotations et leur contexte clinique sur 50 PDS (20 distincts pour chaque annotateur et 10 communs)

3. <https://brat.nlplab.org/>

Correction des annotations

1

ANTECEDENTS FAMILIAUX

Relation(FamilyMember) | Trouble(History|FamilyMember) | Evénement(History|FamilyMember)
 mère , dépendante à l'alcool , décédée d'une cirrhose du foie.

Relation(FamilyMember|False|Recent) | Trouble(History|Patient|False|Recent) | Evénement(Patient|False|Recent) | Substance(Patient|False|False|Recent) | Dose
 frère , suicide par pendaison en 3231 dans un contexte de dépression traitement habituel , teralithe lp 400mg/ 2.5 cp le soir zypadhera300mg/inj ,

Evénement(Patient|False|Recent) | TIME[QUANTITE]
 une injection tous les quatorze jours.

2

ANTECEDENTS FAMILIAUX

Relation(History|Patient) | Trouble(History|Patient) | Evénement(Patient|False|Recent) | Substance(Patient|False|False|Recent)
 mère , dépendante à l'alcool , décédée d'une cirrhose du foie.

Relation(FamilyMember|False|Recent) | Evénement(Patient|False|Recent) | TIME[DATE] | Trouble(History|FamilyMember) | Substance(Patient|False|False|Recent) | Dose | Dose
 frère , suicide par pendaison en 3231 dans un contexte de dépression traitement habituel , teralithe lp 400mg/ 2.5 cp le soir zypadhera300mg/inj , une injection tous les quatorze jours.

Evénement(Recent|Patient) | Frequency




Figure 3 – Correction de l'annotation du Pipeline. La partie haute montre la sortie du pipeline et la partie basse, les modifications du correcteur. On constate, par exemple, que l'annotation de « une injection tous les quatorze jours » indique une durée là où il fallait voir une fréquence.

Évaluation des annotations

Table 1 – Résultats quantitatifs des évaluations de l'extraction d'entités nommées par les 2 annotateurs avec un accord inter annotateur de 0.88.

		Quantity	Precision	Recall	F1
Sign Or Symptom		1747	0.9544	0.9503	0.9524
Disease		150	0.9826	0.7635	0.8593
Trouble		459	0.9894	0.9493	0.9690
Clinical Event		1459	0.9744	0.8140	0.8870
Personal Situation		188	0.9895	0.8468	0.9126
Drug	Drug Name	1034	0.8200	0.9805	0.8822
	Drug Dose	650	0.9848	0.9610	0.9727
Temporal Inf.	Date	840	0.9942	0.9709	0.9824
	Duration	529	0.9574	0.9777	0.9674
	Frequency	212	0.9459	0.8373	0.8883

Évaluation des annotations

Table 2 – Résultats quantitatifs des évaluations de l'extraction d'entités nommées par les 2 annotateurs avec un accord inter annotateur de 0.88.

		Quantity	Precision	Recall	F1
Sign Or Symptom		1747	0.9544	0.9503	0.9524
Disease		150	0.9826	0.7635	0.8593
Trouble		459	0.9894	0.9493	0.9690
Clinical Event		1459	0.9744	0.8140	0.8870
Personal Situation		188	0.9895	0.8468	0.9126
Drug	Drug Name	1034	0.8200	0.9805	0.8822
	Drug Dose	650	0.9848	0.9610	0.9727
Temporal Inf.	Date	840	0.9942	0.9709	0.9824
	Duration	529	0.9574	0.9777	0.9674
	Frequency	212	0.9459	0.8373	0.8883

Conclusion

Résumé

L'objectif de ce travail est de reconstituer les données structurées des patients à partir du CRH afin de compléter la chronologie des patients dans le projet PsyCARE.

À partir des textes narratifs psychiatriques français, nous avons pu :

- Effectuer une annotation sémantique en utilisant les plugins GATE et les algorithmes que nous avons modifiés pour les adapter à la structure des CRHs
- Évaluer les annotations, révélant que le système est capable d'identifier correctement les concepts de l'ontologie

Limitations

Limitations

Malgré les bons résultats obtenus, l'approche présente les limites suivantes :

- Les syntagmes nominaux obtenus automatiquement ne sont pas toujours parfaits et peuvent donner lieu à des erreurs
- La taille de l'ensemble de données annotées reste limitée dans ce processus d'évaluation par rapport à d'autres travaux
- Le système est principalement basé sur des techniques traditionnelles de traitement des langues et, pour l'améliorer, des règles doivent être ajoutées manuellement

Discussion

Pour pallier les limites de notre approche :

- Une combinaison d'approches basées sur des règles et d'apprentissage automatique peut être utilisée pour construire un système d'annotation sémantique plus efficace et plus générique (Landolsi, Hlaoua et Ben Romdhane 2023)
- L'inclusion des événements cliniques non psychiatriques les plus courants et des maladies dans l'ontologie peut aider à élargir la portée de l'annotation

Perspectives

- Élargir la portée de l'annotation : Identifier les relations entre les concepts (Nasar, Jaffry et Malik 2021)






Remerciements

Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du Programme d'Investissements d'Avenir portant la référence PsyCARE ANR-18-RHUS- 0014.



Références

References I

-  Cunningham, Hamish et al. (2014). “GATE : A Framework and Graphical Development Environment for Robust NLP Tools and Applications.”. In : doi : 10.3115/1073083.1073112.
-  Hilbey, Jacques, Xavier Aimé et Jean Charlet (mai 2022). “Temporal Medical Knowledge Representation Using Ontologies”. In.
-  Landolsi, Mohamed Yassine, Lobna Hlaoua et Lotfi Ben Romdhane (fév. 2023). “Information extraction from electronic medical documents : state of the art and future research directions”. In : *Knowledge and Information Systems* 65.2, p. 463-516. issn : 0219-3116. doi : 10.1007/s10115-022-01779-1.
-  Nasar, Zara, Syed Waqar Jaffry et Muhammad Malik (2021). “Named Entity Recognition and Relation Extraction : State of the Art”. In.
-  Neves, Mariana et Jurica Ševa (déc. 2019). “An extensive review of tools for manual annotation of documents”. In : *Briefings in Bioinformatics* 22.1, p. 146-163. doi : 10.1093/bib/bbz130.