La pré-annotation automatique de textes cliniques comme support au dialogue avec les experts du domaine lors de la mise au point d'un schéma d'annotation

Virgile Barthet, Marie José Aroulanda, Laura Monceaux-Cachard, Christine Jacquin, Cyril Grouin, Johann Gutton, Guillaume Hocquet, Pascal de Groote, Michel Komajda, Emmanuel Morin, et al.

▶ **To cite this version:**

Virgile Barthet, Marie José Aroulanda, Laura Monceaux-Cachard, Christine Jacquin, Cyril Grouin, et al.. La pré-annotation automatique de textes cliniques comme support au dialogue avec les experts du domaine lors de la mise au point d'un schéma d'annotation. Atelier ARTS, 18e Conférence en Recherche d'Information et Applications – 16e Rencontres Jeunes Chercheurs en RI – 30e Conférence sur le Traitement Automatique des Langues Naturelles – 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, Jun 2023, Paris, France. pp.1-7. hal-04131600

HAL Id: hal-04131600
https://hal.science/hal-04131600v1

Submitted on 20 Jun 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.

Distributed under a Creative Commons Attribution 4.0 International License

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

La pré-annotation automatique de textes cliniques comme support au dialogue avec les experts du domaine lors de la mise au point d'un schéma d'annotation

Virgile Barthet[1] Marie José Aroulanda[2] Laura Monceaux-Cachard[3]
Christine Jacquin[3] Cyril Grouin[1] Johann Gutton[2] Guillaume Hoquet[2]
Pascal de Groote[4] Michel Komajda[2] Emmanuel Morin[3] Pierre Zweigenbaum[1]
(1) Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, Orsay, France
(2) Hôpital Saint Joseph, DIMID et Service de Cardiologie, Paris, France
(3) Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, Nantes, France
(4) CHU de Lille, Service de Cardiologie, Lille, France
{prenom.nom}@lisn.fr {laura.monceaux,jacquin-c,emmanuel.morin}@ls2n.fr
{maroulanda,jgutton,ghoquet,mkomajda@ghpsj.fr} pascal.degroote@chu-lille.fr

RÉSUMÉ

La pré-annotation automatique de textes est une tâche essentielle qui peut faciliter l'annotation d'un corpus de textes. Dans le contexte de la cardiologie, l'annotation est une tâche complexe qui nécessite des connaissances approfondies dans le domaine et une expérience pratique dans le métier. Pré-annoter les textes vise à diminuer le temps de sollicitation des experts, facilitant leur concentration sur les aspects plus critiques de l'annotation. Nous rapportons ici une expérience de pré-annotation de textes cliniques en cardiologie : nous présentons ses modalités et les observations que nous en retirons sur l'interaction avec les experts du domaine et la mise au point du schéma d'annotation.

ABSTRACT

Automatic pre-annotation of clinical texts to support the dialogue with domain experts during the design of an annotation schema

Automatic text pre-annotation is an essential task that can facilitate the annotation of a text corpus. In the context of cardiology, manual text annotation is a complex task that requires in-depth domain knowledge and practical professional experience. Pre-annotating texts aims to reduce the time spent by experts on manual annotation and to focus their intervention on more critical aspects of annotation. We report here a pre-annotation experiment for clinical texts in cardiology : we present its modalities and the lessons learnt about our interaction with domain experts and on annotation schema design.

MOTS-CLÉS : TAL, Médical, Cardiologie, Annotation, Pré-annotation, Schéma d'annotation.

KEYWORDS: NLP, Medical, Cardiology, Annotation, Pre-annotation, Annotation schema.

1 Introduction

Selon une étude de l'OMS datant de 2020 (World Health Organization, 2020), les maladies cardiovasculaires sont l'une des principales causes de décès dans le monde et l'analyse des données cliniques joue un rôle crucial dans l'amélioration de la prise en charge des patients. Dans ce contexte,

Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

nous cherchons à analyser les textes de dossiers de patients en cardiologie dans le but de déterminer précocément si un patient insuffisant cardiaque présente un risque de décès dans les trois mois suivant son hospitalisation. Ces textes ont en commun avec les articles scientifiques de traiter d'un domaine spécialisé, et mettent en jeu des entités de même nature. Les méthodes à l'état de l'art pour la détection automatique de ces entités reposent sur l'entraînement de classifieurs supervisés à partir de corpus annotés manuellement. (Patel *et al.*, 2018)

Cependant, l'annotation manuelle de grands volumes de données textuelles cliniques dans le contexte de la cardiologie peut être une tâche complexe et chronophage, nécessitant une connaissance approfondie du domaine et une expérience pratique dans le métier. C'est pourquoi une pré-annotation automatique des textes est souvent mise en place pour faciliter l'annotation humaine d'un corpus de textes et aider à réduire le temps nécessaire à l'annotation, tout en améliorant potentiellement la cohérence de ces annotations (Lingren *et al.*, 2012).

Dans ce contexte, la présente étude se focalise sur les méthodes et ressources utilisées pour pré-annoter des textes cliniques en cardiologie, ainsi que sur l'interaction avec des experts du domaine pour définir et redéfinir les types d'entité, les points d'arbitrage et les méthodes de représentation de l'information. L'objectif est de mettre en évidence l'importance de ces éléments pour l'annotation efficace et précise de données textuelles en cardiologie. L'annotation requiert également d'établir un bon schéma d'annotation, cohérent et qui reflète correctement la manière de penser des experts (Shinohara *et al.*, 2022). Cette étude constitue un retour d'expérience sur la pré-annotation de textes cliniques en cardiologie et les échanges avec des experts non-initiés au domaine du traitement automatique des langues, mettant en évidence les défis spécifiques liés à la communication entre des experts d'un autre domaine et des informaticiens, ainsi que des pratiques pour surmonter ces obstacles. Nous présentons nos méthodes de pré-annotation et les types d'entités de notre schéma d'annotation (section 2), puis les résultats obtenus et les enseignements que nous en tirons (section 4).

2 Pré-annotation automatique des entités

Pour la pré-annotation automatique des entités, nous utilisons des méthodes traditionnelles de détection d'entités. Ces méthodes ont l'avantage d'être agiles, dans la mesure où il est facile et rapide de prendre en compte de nouvelles expressions. Nous les avons utilisées sur un échantillon de documents pendant la mise au point du schéma d'annotation. Leur capacité de généralisation reste néanmoins limitée, d'où l'intérêt de passer à des méthodes par apprentissage supervisé dans une seconde phase une fois le corpus entier annoté et corrigé manuellement.

1. Appariement exact de termes. Nous avons pour cela construit des lexiques à partir de terminologies médicales en français, notamment présentes dans le Metathesaurus de l'UMLS (Bodenreider, 2004), ou proposées par des agences françaises comme la BDPM [1] pour les traitements médicaux. Nous avons aussi créé des listes de termes ne provenant pas de thésaurus préexistants, par exemple dans le cas des entités de type Entourage pour détecter les termes du champ lexical de la famille.
2. Détection de 20 préfixes et 7 suffixes révélateurs. Par exemple, *postero-* et *antero-* suivis de termes ayant le type Anatomie ajoutent un degré de précision supplémentaire sur la localisation anatomique. Selon le contexte, les préfixes *hypo-* et *hyper-* peuvent indiquer des pathologies ou des signes ; le suffixe *-pathie* indique une pathologie, *-émie* indique un paramètre mesurable, et

1. https://base-donnees-publique.medicaments.gouv.fr/

le suffixe *-graphie* peut indiquer un examen d'imagerie. Des exceptions peuvent être ajoutées dans certains cas, par exemple le terme anémie désigne une pathologie et non un paramètre mesurable, malgré la présence du suffixe *-émie*.

3. Plus de 100 mots-clefs, par exemple *non*, *ni* ou *sans* indiquent généralement des négations, et *Hôpital* et *en* des lieux. Certains mots-clefs nécessitent un traitement supplémentaire pour comprendre le contexte de leur utilisation. Par exemple le mot *majoration* aura un sens différent lorsqu'il est utilisé avant un signe, auquel cas il s'agit d'une Dégradation de l'état du patient, ou avant un traitement médical, où il s'agira d'une Augmentation du traitement — ne signifiant pas nécessairement que l'état général du patient change.
4. Environ 20 mots-déclencheurs indiquent le type d'entité du texte qui suit. Par exemple, *Greffe de* ou *Thérapie* indiquent un traitement non médical, alors que *Maladie de* ou *Syndrome de* indiquent des pathologies, et *Tranfert en* ou *Orienté en* expriment des changements de lieu. La différence principale entre un mot-clef et un mot déclencheur est que le mot-clef se suffit généralement à lui même (le mot *non* suffit à indiquer une négation par exemple), alors que le mot déclencheur annonce d'autres mots qui vont venir compléter le terme. Par exemple *Syndrome* à lui tout seul ne donne pas d'information pertinente sur l'état du patient, alors que *Syndrome coronarien aigu* si.
5. 25 expressions régulières sont utiles pour traiter les diverses valeurs numériques comme les Dates ou les Valeurs composées de chiffres (par exemple : 42cm², 2020-06-12 etc.).

3 Types d'entités et évolutions

Nous décrivons maintenant les principaux types d'entités de notre schéma d'annotation. Ces types d'entités sont le résultat de discussions entre TAListes et cliniciens. La table 2 de l'annexe A donne la liste complète de ces types d'entités.

Signes et symptômes, Pathologies Les signes et symptômes sont utilisés pour apprécier l'état de santé du patient et déterminer ses pathologies. Dans notre projet, la pathologie clé est l'insuffisance cardiaque qui est responsable de symptômes tels que l'essoufflement ou la fatigue anormale et qui entraîne l'apparition de signes cliniques comme des crépitants. Les pathologies incluent également les comorbidités, qui peuvent être observées simultanément à l'insuffisance cardiaque et apporter un risque supplémentaire.

Initialement, ces entités formaient une seule classe. Après examen de l'annotation résultante, les cliniciens ont fait valoir l'intérêt de distinguer les trois sous-classes Signes, Symptômes, et Pathologies. La mise à jour des ressources de pré-annotation a montré la difficulté à distinguer clairement Signes et Symptômes, pointant des exemples où une même observation apparaissait comme l'un ou l'autre selon le point de vue. Cela a abouti a un accord sur deux classes : Signes et symptômes, et Pathologies.

Examens et traitements Les examens représentent les différents tests, procédures et analyses médicales effectuées sur le patient pour diagnostiquer ou évaluer sa condition médicale, tels que l'échocardiographie ou l'électrocardiogramme. Les traitements représentent les différentes options de prise en charge thérapeutique, comme les diurétiques ou les béta-bloquants, ainsi que les interventions chirurgicales ou la pose de dispositifs médicaux tels que les stents.

Le schéma initial prévoyait des paramètres mesurables comme le poids ou la FEVG (fraction d'éjection du ventricule gauche). La pré-annotation de ces entités a fait ressortir la non-annotation des examens eux-mêmes, comme l'ECG (électrocardiogramme). Après discussion, ils ont été ajoutés au schéma d'annotation et les ressources de pré-annotation ont été mises à jour.

Vie du patient Il s'agit d'un ensemble de variables qui décrivent les aspects socio-démographiques et comportementaux du patient, ainsi que son environnement social. Les comportements du patient comprennent des informations sur son mode de vie et ses habitudes, comme l'activité physique, le régime alimentaire, la consommation de tabac et d'alcool, etc. L'autonomie du patient fait référence à sa capacité à réaliser les activités quotidiennes de manière indépendante, ainsi qu'à ses besoins en matière de soutien et d'assistance. L'entourage du patient comprend des informations sur les personnes qui gravitent autour du patient, comme les membres de sa famille ou les aidants, et leur rôle dans la prise en charge du patient. Les entités Vie du patient peuvent fournir des informations importantes pour comprendre le contexte dans lequel se déroule la prise en charge de l'insuffisance cardiaque, et permettent d'évaluer les facteurs de risque et les déterminants sociaux de la santé qui peuvent influencer l'évolution de la maladie.

À ce stade, la pré-annotation reste incomplète et montre une difficulté particulière à cerner les éléments qui concourent à l'évaluation de l'autonomie et de l'entourage. Cela nous a amenés à démarrer un sous-projet spécifique pour mieux les déterminer.

Entités auxiliaires permettant d'annoter des informations, généralement pour préciser des entités de type Signe, Examen, Traitement, etc. :

- temporelles (date, âge du patient, durée d'un traitement...) ;
- localisation du patient (lieu de séjour) et ses éventuels changements (provenance, destination) ;
- localisation anatomique et valeurs précisant les signes ou traitements ;
- évolution ou absence de signes ou de traitements.

Ici encore, la pré-annotation nous a donné un support concret pour des allers-retours avec les experts du domaine. Cela a mené notamment à étendre le champ d'application des valeurs, initialement uniquement associées aux paramètres mesurables, aux signes et symptômes.

La plupart des types d'entités présentés ici se rencontrent dans les articles scientifiques du même domaine. Par exemple, des types d'entités tels que *pathologie*, *anatomie*, *paramètres mesurables*, *traitement (non) médical* et *examens* peuvent être pertinents pour l'annotation des articles scientifiques dans des domaines tels que la médecine et la biologie. Étant donné que la même expertise serait mobilisée, le travail présenté ici devrait être représentatif au moins d'une partie du travail qui devrait être fait pour l'extraction d'informations de textes scientifiques.

4 Résultats et enseignements tirés

Nous avons évalué la qualité de la pré-annotation sur 3 textes entièrement annotés par un expert du domaine. Nous avons pour cela utilisé l'outil brat-eval [2] qui évalue la correction des types et des frontières d'entités. Le mode d'évaluation utilisé pour les types est *EXACT* : brat-eval indique que le type est correct uniquement si le type d'entité annoté par l'expert et celui produit par le programme

2. https://github.com/READ-BioMed/brateval

sont identiques. Le mode de respect des frontières est *OVERLAP*, ce qui donne une flexibilité dans la délimitation des entités : par exemple, étant donné le terme *un carcinome canalaire*, si l'expert annote l'expression entière comme une Pathologie alors que le programme annote seulement *carcinome canalaire* comme Pathologie, brat-eval considère quand même que l'annotation est correcte. Les résultats figurent dans la table 1.

Type	Précision	Rappel	Nb	Type	Précision	Rappel	Nb
Pathologie	0,70	1,00	76	Autonomie	1,00	1,00	4
Signe_Symptome	0,87	0,68	111	Entourage	0,68	1,00	11
Gravité	0,43	0,29	35	Date	0,90	0,29	38
Hypothétique	0,80	0,67	8	Fréquence	1,00	0,94	34
Examen	0,94	0,88	26	Âge	1,00	1,00	7
Param_mesurable	0,89	0,78	78	Lieu	1,00	0,57	22
Param_physiologique	1,00	1,00	3	Anatomie	0,58	0,74	86
Trt_non_médical	0,91	0,83	25	Valeur	0,81	0,75	151
Trt_médical	1,00	0,77	107	Négation	0,57	0,87	76
Comportement	1,00	1,00	6	Évolution	1,00	0,31	46
Heure	0,63	1,00	7	Mode	1,00	0,86	14
Concentration	0,94	0,76	42	Chgt_lieu	0,50	0,55	11

TABLE 1 – Évaluation de la pré-annotation selon brat-eval, en précisant le mode d'évaluation *EXACT* et le mode de respect des frontières *OVERLAP*.

Le travail de pré-annotation a servi de support pédagogique auprès des experts non informaticiens. En effet, certains concepts de traitement automatique des langues nécessitent du temps pour être assimilés correctement, tandis que certains choix peuvent paraître obscurs et nécessiter des explications. Montrer une annotation automatique sur l'ensemble des textes en cours d'examen a permis d'illustrer les points discutés et de faire réagir les cliniciens sur l'état courant de la modélisation. Cela s'est avéré utile lors des échanges réguliers avec les cliniciens pour mieux comprendre leur point de vue sur la manière dont certaines choses sont annotées.

Nous avons constaté que la pré-annotation a été globalement bien reçue par les cliniciens, qui ont compris très clairement l'intérêt de ce travail. Les avantages en termes de gain de temps et de réduction de la charge mentale ont été ressentis. Les résultats de la table 1 montrent capacité de la pré-annotation à produire des résultats de qualité suffisamment proche de l'annotation manuelle des experts pour que l'on puisse estimer que leur temps de correction reste inférieur à une annotation manuelle ex nihilo.

La pré-annotation a également été un outil précieux pour tester et valider le contenu du schéma d'annotation en cours de développement. Les discussions autour des choix de pré-annotation ont permis d'identifier des incohérences ou des manques dans le schéma d'annotation initial, qui ont été corrigés en conséquence. De plus, la pré-annotation a permis de mettre à l'épreuve le schéma d'annotation en identifiant les cas d'utilisation courants et les cas plus complexes, ce qui a permis de renforcer et d'affiner ce schéma de manière itérative. Nous notons enfin que des méthodes et outils ont été proposés pour opérationnaliser les opérations de pré-annotation que nous avons décrites plus haut (Lison *et al.*, 2021), que nous envisageons d'utiliser dans le futur.

Remerciements

Ce travail a été soutenu par l'Agence Nationale pour la Recherche (ANR) dans le cadre du projet PREDHIC (ANR-21-CE23-0039).

Références

BODENREIDER O. (2004). The Unified Medical Language System (UMLS) : Integrating biomedical terminology. *Nucleic Acids Research*, **32**(Database issue), D267–270.

LINGREN T., DELEGER L., MOLNAR K., ZHAI H., MEINZEN-DERR J., KAISER M., STOUTENBOROUGH L., LI Q. & SOLTI I. (2012). Pre-annotating clinical notes and clinical trial announcements for gold standard corpus development : Evaluating the impact on annotation speed and potential bias. In *2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology*, p. 108–108. DOI : 10.1109/HISB.2012.33.

LISON P., BARNES J. & HUBIN A. (2021). skweak : Weak supervision made easy for NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing : System Demonstrations*, p. 337–346, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-demo.40.

PATEL P., DAVEY D., PANCHAL V. & PATHAK P. (2018). Annotation of a large clinical entity corpus. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 2033–2042, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1228.

SHINOHARA E., SHIBATA D. & KAWAZOE Y. (2022). Development of comprehensive annotation criteria for patients' states from clinical texts. *J Biomed Inform*, **134**, 104200.

WORLD HEALTH ORGANIZATION (2020). Leading causes of death and disability — a visual summary of global and regional trends 2000-2019. consulté le 31/3/2023.

A Types d'entités

Type	Description	Type	Description
Pathologie	Maladie responsable de symptômes et signes cliniques, y compris comorbidité.	Signe et Symptôme	Signe : manifestation d'une maladie, constatée par un observateur. Symptôme : signe dont le malade se plaint.
Gravité	Sévérité d'un signe ou d'une pathologie.	Hypothétique	Incertitude sur la présence d'un signe ou d'une pathologie.
Examen	Examen (d'imagerie, etc.).	Paramètre mesurable	Variable mesurée par un examen.
Paramètre physiologique	Variable désignant une fonction normale (physiologique) du patient.	Traitement médical	Traitement médicamenteux, typiquement nom de médicament.
Traitement non médical	Traitement autre que médicament.	Mode d'administration	Mode d'administration d'un médicament ou lié à un paramètre mesurable.
Comportement	Comportement du patient.	Autonomie	Information sur le degré d'autonomie du patient.
Entourage	Information sur la présence de famille, etc. dans l'entourage du patient.	Date	Date d'un événement, absolue (chiffrée) ou relative (non chiffrée).
Heure	Heure d'un événement, absolue (chiffrée) ou relative (non chiffrée).	Durée	Durée d'un événement, absolue ou relative.
Fréquence	Fréquence d'un traitement ou d'un comportement.	Âge	Âge du patient.
Lieu	Lieux de provenance, de séjour et de sortie du patient.	Changement de lieu	Changements de lieux qui pourraient indiquer une évolution de l'état du patient.
Anatomie	Parties du corps.	Valeur	Valeur qualitative (textuelle) ou quantitative (numérique) d'une entité.
Concentration	Concentration d'un médicament.	Négation	Absence d'entité.
Évolution	Évolution d'un signe, un traitement ou de l'état général du patient.	Dose	Dosage d'un traitement médical.

TABLE 2 – Liste des types d'entités