Extraction de relations dans des comptes-rendus d'infectiologie : expériences sur le corpus PARHAF

Thomas Checchin[1] Julien Jacques[1] Adrien Guille[1]
(1) Université Lumière Lyon 2, Université Claude Bernard Lyon 1, ERIC, 69007, Lyon, France
thomas.checchin@univ-lyon2.fr, julien.jacques@univ-lyon2.fr,
adrien.guille@univ-lyon2.fr

RÉSUMÉ

La tâche d'extraction de relations est fortement dépendante de la syntaxe. Le codage d'une phrase selon le formalisme AMR, une représentation sémantique abstraite sous forme de graphe, décrit les liens entre les concepts sémantiques de la phrase. Notre travail explore l'extraction de relations cliniques intra-phrases, sur la base des codages AMR, à l'aide de classifieurs simples. Nous nous appuyons sur un nouveau corpus français de rapports cliniques en infectiologie PARHAF ainsi qu'un analyseur AMR pré-existant.

ABSTRACT

Relationship extraction from reports on infectiology cases : experiments on the PARHAF corpus

The task of relation extraction is highly syntax-dependent. Coding a sentence into the AMR framework, an abstract semantic representation in graph form, establishes the links between the semantic concepts in a sentence. Our work explores the extraction of intra-sentence clinical relations, based on AMR encodings, using simple classifiers. We rely on a new French corpus of clinical reports in infectious diseases, PARHAF, as well as a pre-existing AMR parser.

MOTS-CLÉS : Extraction de relations, AMR, comptes-rendus médicaux.

KEYWORDS: Relationship extraction, AMR, medical reports.

1 Introduction

La discipline du traitement automatique du langage naturel (TALN) appliquée au domaine médical repose sur l'exploitation de comptes-rendus ou bilans de sorties cliniques non structurés. Alors que des corpus anglophones sont disponibles de longue date, *e.g.* MIMIC (Neamatullah *et al.*, 2008), des premiers corpus francophones viennent d'être partagés avec la communauté, à savoir PARROT (Guellec *et al.*, 2025) et PARHAF (Tannier *et al.*, 2026). Dans ce court article, nous étudions une sous-partie de ce dernier, en rapport avec l'infectiologie et qui comporte 7 394 documents, rédigés par des experts en français, portant sur 5 009 patients fictifs dans 20 spécialités médicales différentes. Les entités importantes en infectiologie et les relations qu'elles entretiennent y sont annotées [1]. Nous explorons l'utilisation d'un langage de représentation sémantique abstraite, *Abstract Meaning Representation* ou AMR (Banarescu *et al.*, 2013), dans le cadre de l'extraction de relations appliquée au domaine du médical. Notre travail exploite cette représentation AMR pour factoriser les

1. Le jeu de données est librement accessible sur HuggingFace (https://huggingface.co/datasets/HealthDataHub/PARHAF-infectiology-annotated)

concepts et les rôles afin de tenter d'entraîner des classifieurs simples et interprétables (des arbres de décision ou des régressions logistiques) sans recourir à un apprentissage de modèles à base de plongements lexicaux (Touchent *et al.*, 2024; Labrak *et al.*, 2023).

2 Présentation des représentations AMR

L'AMR (Banarescu *et al.*, 2013) est un formalisme sémantique représentant le sens d'une phrase indépendamment de sa syntaxe, de sorte que différentes phrases ayant la même signification ont la même représentation. Cette représentation AMR est défini par un graphe orienté, enraciné et acyclique. Les nœuds correspondent à des concepts sémantiques intervenant dans la phrase et définis par PropBank (Kingsbury & Palmer, 2002) ou des entités abstraites. Quant aux arcs, ils représentent les relations sémantiques entre ces concepts comme des arguments typés, *:ARGn*, décrivant qui est l'agent ou l'objet d'une action ou des rôles généraux (*e.g.*, « :location», « :time»). Outre sa représentation en graphe, l'AMR s'écrit textuellement via la notation Penman (Kasper, 1989). Ce formalisme capture non seulement des phénomènes complexes tels que les coréférences, les entités nommées ou encore les rôles thématiques, mais aussi leurs interactions tout en se déliant de la syntaxe.

Cette capacité à transformer le langage naturel en une représentation normalisée répond aux enjeux du domaine de la santé. En effet, les rapports cliniques se caractérisent par des phrases chargées de termes précis, d'abréviations et de nuances critiques (négations, temporalités, incertitudes). Dans le but d'extraire les informations clefs du diagnostic médical, l'AMR semble une approche intéressante en jouant le rôle de pont entre la complexité du langage naturel et la nécessité de structurer les documents médicaux. Notre travail est ainsi motivé par ces représentations qui mettent en évidence les relations sémantiques pour organiser les énoncés ; les relations cliniques entre les concepts sont alors, a priori, dégagées dans ces représentations. Pour convertir un texte en une représentation AMR au format textuel Penman, nous appliquons un analyseur AMR pour le français, GemmAMR[2] (Checchin *et al.*, 2026).

3 Expériences sur le sous-corpus PARHAF

Description du sous-corpus d'infectiologie Le sous-corpus d'infectiologie est constitué de 134 documents regroupant 5 289 annotations et 1 713 relations. Les données sont scindées en deux ensembles pour l'entraînement (80 %) et la validation (20 %). L'annotation repose sur quatre types de concepts (**infection**, **site**, **bactériémie**, **bactérie**), et la présence, l'absence ou l'incertitude de ceux-ci. Les entités forment trois relations : **Agent Pathogène** (497), **Origine** (685) et **Site Primaire** (492).

Parmi les 13 051 phrases extraites de ces documents via la bibliothèque NLTK (Bird *et al.*, 2009), les relations apparaissent dans seulement 636 phrases distinctes (une phrase pouvant apparaître à l'identique dans plusieurs documents). Il est à noter que 97 % des relations relient des entités apparaissant à l'intérieur de la même phrase. Comme le formalisme AMR est optimisé pour l'échelle de la phrase et la majorité de l'information y est concentrée, nous choisissons alors de restreindre nos expériences à ces portions de documents contenant ces relations.

2. Disponible en libre accès sur HuggingFace (https://huggingface.co/AdrienGuille/GemmAMR-fr-v1)

Méthode	Exactitude			Macro F-mesure		
	Phrase	AMR	Étiquettes extraites	Phrase	AMR	Étiquettes extraites
Arbre de décision	94 %	96 %	96 %	93 %	95 %	95 %
Régression logistique	97 %	97 %	96 %	97 %	97 %	96 %

TABLE 1 – Résultats de l'exactitude et de la macro F-mesure pour la classification des relations

Méthode	Phrase			AMR			Étiquettes extraites		
	A	O	S	A	O	S	A	O	S
Arbre de décision	90 %	100 %	90 %	93 %	100 %	93 %	93 %	100 %	93 %
Régression logistique	96 %	100 %	96 %	96 %	99 %	96 %	94 %	100 %	94 %

TABLE 2 – Résultats de la F-mesure pour la classification de chaque relation (A désigne la relation **Agent Pathogène**, O la relation **Orgine** et S la relation **Site Primaire**)

Classification des relations Pour permettre une classification simple et interprétable, nous utilisons des estimateurs tels que les arbres de décision ou les régressions logistiques. Grâce à l'analyseur AMR, les représentations sont générées sur les phrases du corpus contenant des relations. Du fait que ce sont des graphes, les concepts et les rôles y peuvent être extraits pour supprimer notamment ceux sans valeurs ajoutées à la classification comme *:ARGn*. L'entrée des modèles correspond à un triplet contenant le type de séquences (la phrase, l'AMR ou les étiquettes extraites du graphe) et les deux entités préalablement vectorisées par la fréquence d'occurrences de leurs jetons. Les résultats pour ces modèles et ces types d'entrées sont comparables et diffèrent selon le modèle (*cf.* tables 1 et 2). Toutefois, nous avons observé que les représentations basées sur l'AMR ou les étiquettes extraites des graphes ont une plus grande stabilité vis-à-vis des variations d'hyperparamètres et des méthodes de vectorisation utilisées, qu'il s'agisse de la fréquence brute des jetons ou de TF-IDF. L'optimisation des hyperparamètres a été effectuée par une recherche par grille. L'exploration des arbres de décisions a concerné le critère de scission et la pondération des relations. Pour la régression logistique, les paramètres ont été choisis selon le solveur, le terme de régularisation et le ratio de la norme L_1. Au vu des résultats obtenus pour les différentes méthodes et les types d'entrée, la tâche de classification des relations étant donné les entités apparaît aisée.

Filtrage et classification binaire selon les entités La difficulté résidant dans la première étape de l'extraction d'information, *i.e.* la détection d'entités, nous tentons d'aborder le problème sous l'angle d'une classification binaire : déterminer si une phrase contient ou non une relation entre les entités identifiées. Pour ce faire, les entités du jeu d'entraînement sont utilisées comme filtres sur l'échantillon de validation afin de sélectionner sans apprentissage les phrases candidates à une relation. Un nouveau triplet est ensuite formé pour chaque instance filtrée, à l'instar de l'expérience précédente. Cependant, les résultats indiquent une forte proportion de faux positifs : de nombreux triplets sont extraits, mais peu d'entre eux correspondent à des relations réelles. Par conséquent, l'estimateur peine à généraliser et surestime la présence de relations alors qu'elles sont minoritaires. Par ailleurs, le filtrage ne permet que de capturer un tiers des entités, puisque la syntaxe des entités est légèrement différente et certaines entités sont très longues (jusqu'à 10 mots pour une entité).

4 Conclusion et perspectives

Nous avons étudié l'utilisation d'un analyseur AMR dans un cadre clinique, sans ajustement sur des documents du domaine de la santé. Nous comparons les graphes résultants de l'analyseur et les phrases pour la classification de relations. Notre analyse s'est appuyée sur le nouveau corpus français PARHAF annoté pour l'extraction d'information en infectiologie. Nous observons que les relations se classifient simplement une fois l'extraction d'entités réalisée que ce soit avec ou sans le pré-traitement AMR. De plus, nous observons que la difficulté de ce sous-corpus repose sur l'extraction d'entités de termes précis du domaine médical tels que les zones ou les étapes d'une infection (plus susceptibles de dépendre de la syntaxe). Les résultats indiquent qu'un modèle encodeur basé sur BERT n'est pas justifié sur la tâche de classification de relations. En revanche, son utilisation serait pertinente pour la détection d'entités. Une piste pour poursuivre ce travail avec ce formalisme serait de rendre l'analyseur plus robuste à la syntaxe et au domaine médical, notamment par un apprentissage de la sémantique de textes cliniques, dans l'espoir d'obtenir une détection et une représentation plus consistante des entités médicales dans les AMR. Une autre approche serait de remplacer les classifieurs simples par un GNN minimaliste, bénéficiant de la standardisation AMR, et exploitant la structure du graphe.

Références

BANARESCU L., BONIAL C., CAI S., GEORGESCU M., GRIFFITT K., HERMJAKOB U., KNIGHT K., KOEHN P., PALMER M. & SCHNEIDER N. (2013). Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*.

BIRD S., KLEIN E. & LOPER E. (2009). *Natural Language Processing with Python*. " O'Reilly Media, Inc.".

CHECCHIN T., JACQUES J. & GUILLE A. (2026). Un décodeur pour l'analyse sémantique AMR en français. In *33e Conf'erence sur le Traitement Automatique des Langues Naturelles*.

GUELLEC B. L., ADAMBOUNOU K. & AL. (2025). Parrot, an open multilingual radiology reports dataset. *European Journal of Radiology Artificial Intelligence*.

KASPER R. (1989). A flexible interface for linking applications to penman's sentence generator. In *Speech and Natural Language : Proceedings of a Workshop Held at Philadelphia, Pennsylvania*.

KINGSBURY P. & PALMER M. (2002). From TreeBank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*.

LABRAK Y., BAZOGE A., DUFOUR R., ROUVIER M., MORIN E., DAILLE B. & GOURRAUD P.-A. (2023). Drbert : A robust pre-trained model in french for biomedical and clinical domains.

NEAMATULLAH I., DOUGLASS M. M., LEHMAN L.-W. H., REISNER A., VILLARROEL M., LONG W. J., SZOLOVITS P., MOODY G. B., MARK R. G. & CLIFFORD G. D. (2008). Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, **8**(1), 32. DOI : 10.1186/1472-6947-8-32.

TANNIER X., ABBARA S., FLICOTEAUX R., KHALIL Y., NÉVÉOL A., ZWEIGENBAUM P. & BACRY E. (2026). Parhaf, a human-authored corpus of clinical reports for fictitious patients in french.

TOUCHENT R., ROMARY L. & DE LA CLERGERIE E. (2024). Camembert-bio : Leveraging continual pre-training for cost-effective models on french biomedical data.