Équilibrer qualité et quantité : comparaison de stratégies d'annotation pour la reconnaissance d'entités nommées en cardiologie

V. Barthet[1], M.-J. Aroulanda[3], L. Monceaux-Cachard[2], C. Jacquin[2], C. Grouin[1], J. Weller[3], P. de Groote[4], G. Hocquet[3], M. Komajda[3], E. Morin[2], P. Zweigenbaum[1]

[1]Université Paris-Saclay, CNRS, LISN, Orsay, France
[2]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
[3]Hôpital Saint Joseph, DIMID, et Service de Cardiologie, Paris, France
[4]CHU de Lille, Service de Cardiologie, Lille, France

{virgile.barthet,cyril.grouin,pierre.zweigenbaum}@universite-paris-saclay.fr,
{laura.monceaux,jacquin-c,emmanuel.morin}@univ-nantes.fr, maroulanda,jweller,ghocquet,mkomajda@ghpsj.fr,
pascal.degroote@chu-lille.fr

Résumé

Cet article explore l'entraînement de modèles de reconnaissance d'entités nommées pour l'analyse de dossiers médicaux électroniques. Face au manque d'annotations de qualité, nous étudions trois types d'annotations, expert, non-expert et pré-annotation, et leurs combinaisons. Nous cherchons la combinaison optimale d'annotations menant à l'entraînement d'un modèle transformeur performant. Nous constatons que de nombreux textes pré-annotés mènent à un modèle plus performant qu'une quantité limitée de textes pré-annotés, puis corrigés par un expert, et que l'ajout d'une quantité modérée de textes pré-annotés puis corrigés par un non-expert augmente encore cette performance.

Mots-clés

Traitement automatique des langues, Reconnaissance d'entités, Domaine médical, Pré-annotation, Apprentissage automatique, Cardiologie

Abstract

This article explores training named entity recognition models to analyze electronic health records. Faced with a lack of high-quality annotations, we study three types of annotations, expert, non-expert, and pre-annotation, and their combinations. We look for an optimal combination of annotations that leads to training a high-performing Transformer model. We observe that a large number of pre-annotated texts leads to a better-performing model than a limited number of pre-annotated then expert-reviewed texts, and that adding a moderate number of pre-annotated then non-expert-reviewed texts further increases this performance.

Keywords

Natural Language Processing, Entity Recognition, Medical Domain, Pre-annotation, Machine learning, Cardiology

1 Introduction

La reconnaissance d'entités nommées constitue l'une des tâches fréquentes en traitement automatique des langues (TAL). Elle consiste à détecter dans un texte les informations correspondant à des types d'entités fixés au préalable selon un schéma d'annotation. Ici, nous explorons le contenu de dossiers médicaux électroniques, qui contiennent une richesse de données textuelles pouvant être exploitées pour améliorer la prise en charge des patients. Les types d'entités y sont entre autres les signes et symptômes, maladies, traitements relevés chez un patient.

L'un des principaux défis rencontrés dans le domaine médical réside dans le manque d'annotations de qualité pour entraîner et évaluer les modèles de reconnaissance d'entités nommées (REN). Comme les données annotées sont plus nombreuses en anglais, certains travaux explorent des méthodes translingues pour bénéficier des ressources en anglais ou dans d'autres langues [14, 5]. Par ailleurs, la quantité de textes bruts disponibles est souvent importante, mais le nombre d'annotations manuelles reste limité. De plus, les annotations disponibles peuvent provenir de différentes sources et présenter des niveaux de qualité variables.

Nous abordons cette problématique en nous concentrant sur l'analyse de dossiers de patients dans un service de cardiologie, pour lesquels nous disposons de trois types d'annotations : celles réalisées par des experts du domaine médical, celles effectuées par des non-experts, et celles générées automatiquement par des méthodes de pré-annotation. La pré-annotation automatique peut être appliquée à tous les textes disponibles, mais sa qualité n'est pas optimale. La correction humaine de cette pré-annotation améliore sa qualité, et ce d'autant plus que les annotateurs sont experts du domaine. Mais comme cette correction est chronophage et que les experts sont peu disponibles, la quantité de textes corrigés disponibles décroît avec l'expertise des annotateurs.

De plus, les annotations expertes ne sont pas toujours issues d'un consensus en raison de la difficulté à trouver des annotateurs experts et volontaires pour investir le temps nécessaire à une annotation de grande qualité.

Nous explorons ici la question suivante : comment optimiser les performances d'un modèle de reconnaissance d'entités nommées médicales dans ce contexte en utilisant au mieux les différentes stratégies d'annotation, y compris les données pré-annotées puis corrigées par des experts, les données pré-annotées puis corrigées par des non-experts et les pré-annotations automatiques ?

Dans cet article, nous présentons des travaux antérieurs (sec. 2), la nature des données dont nous disposons (sec. 3.1), le schéma d'annotation employé pour annoter ces données (sec. 3.2). Nous évaluons la contribution de différents groupes de données annotées en mesurant les performances d'un modèle entraîné sur chaque groupe (sec. 4). Nous discutons les résultats obtenus (sec. 5), puis concluons (sec. 6).

2 Travaux antérieurs

Détection d'entités dans des textes cliniques. Les méthodes traditionnelles de détection d'entités dans des textes reposent sur des lexiques et des patrons qui énoncent explicitement quelle forme peuvent prendre les entités de chacun des types définis [6, 11]. Ces méthodes manquent cependant de robustesse face aux variations diverses rencontrées en langue naturelle. Les méthodes d'apprentissage supervisé, que l'on entraîne avec des exemples de textes annotés manuellement, atteignent une meilleure robustesse si suffisamment d'exemples sont disponibles. Les méthodes de l'état de l'art reposent sur des réseaux de neurones de type transformeur : les modèles de langue masqués de la famille BERT [4, 1]. Les grands modèles de langue autorégressifs ont atteint de bonnes performances même sans exemples ou avec peu d'exemples ; cependant, pour la détection d'entités dans le domaine biomédical, ils restent en deçà des modèles de langue masqués de type BERT [7] même lorsque peu d'exemples annotés sont disponibles [13]. De plus, les meilleurs modèles propriétaires comme GPT-4 ne sont pas utilisables sur les données confidentielles des dossiers de patients. Nous utilisons donc ici des modèles de type BERT.

Pré-annotation automatique. Lors de la préparation d'un jeu de données annotées, des méthodes relativement simples de détection d'entités sont souvent appliquées pour fournir une *pré-annotation* des textes qui est ensuite corrigée par les annotateurs humains. Le bénéfice escompté est double : réduire le temps d'annotation et augmenter sa cohérence [10]. C'est ce que nous faisons ici [2].

Apprentissage faiblement supervisé. Lorsque des textes non-annotés sont disponibles en grande quantité, on peut chercher à les inclure dans le processus d'entraînement en trouvant des moyens de les annoter sans intervention humaine, dans un objectif d'apprentissage semi-supervisé [9], ou avec une intervention humaine minimale, dans un cadre d'apprentissage faiblement supervisé [11]. Nous nous plaçons en partie dans ce dernier cadre : nous employons notre méthode de pré-annotation à base de lexiques et de patrons pour créer automatiquement des annotations sur l'ensemble de notre corpus et entraînons un système supervisé sur la base de ces annotations sans correction humaine.

Comme la pré-annotation seule n'est pas parfaite, un système supervisé entraîné sur son résultat risque de reproduire en grande partie ses faiblesses. Nous examinons ici si la constitution d'un corpus d'entraînement combinant des textes pré-annotés automatiquement et des textes corrigés par des annotateurs humains permet de dépasser ces limites.

3 Présentation des données

3.1 Nature des textes

Le travail présenté s'inscrit dans le cadre du projet ANR PREDHIC. Au sein de ce projet, nous avons une collaboration avec les services de cardiologie de deux hôpitaux (le Groupe Hospitalier Paris Saint-Joseph et le CHU de Lille) qui nous permettent de travailler sur des extraits de dossiers de patients sélectionnés et désidentifiés. Chacun des textes que nous appelons par la suite « dossier » est issu d'un séjour d'un patient à l'hôpital Saint-Joseph. Un unique numéro de séjour correspond à la visite à l'hôpital d'un patient, ainsi que son éventuel suivi dans le temps, y compris suite à sa sortie de l'hôpital.

Ces textes sont formés à partir de notes cliniques incluant :

- Compte rendu initial d'examen du patient par un médecin, dressant généralement une liste de symptômes, de pathologies, etc.
- Liste d'examens et leurs résultats
- Liste de traitements médicaux (médicaments) pris par le patient
- Liste de pathologies antérieures, traitées ou non, ainsi que d'antécédents familiaux
- Suivi du patient tout au long du séjour, qui vient compléter ou préciser les informations précédentes, ou qui fait état de changements (arrêt d'un traitement, dégradation de l'état, nouvelles pathologies, changement de la valeur d'un examen, etc.)
- Précisions sur le futur de la prise en charge du patient suite à la fin de son séjour (futurs examens, traitements, etc.)

3.2 Schéma d'annotation

Notre schéma d'annotation repose sur les classes suivantes ; il reprend des notions générales présentes dans d'autres schémas [12, 3, 15] tout en s'adaptant aux besoins spécifiques du projet.

Les entités « *État patient* » regroupent les classes "Pathologie" et "Signe Symptôme" qui décrivent les maladies qui touchent le patient et les symptômes provoqués par ces maladies, ainsi que la classe "Évolution" qui sert à caractériser l'évolution de ces maladies ou de l'état général du patient.

Les entités « *Examen et traitement* » regroupent les examens (scanner, IRM, échocardiographie...), les paramètres mesurables qui sont le plus souvent des grandeurs physiques (taille d'un nodule par exemple), et les traitements médicaux (médicaments, oxygène...), intervention-

nels (ablation, pose d'un stent...) et les traitements qui ne rentrent pas dans ces deux catégories (régimes particuliers, rééducation...). Les traitements peuvent également être complétés d'entités "Dose", "Concentration" et "Mode" (comprimé oral, intraveineux, ampoules...).

Les entités « *Caractéristiques* » complètent les informations données par les autres classes. "Anatomie" précise la partie du corps où se manifeste un symptôme par exemple, "Valeur" permet d'annoter la valeur obtenue par un examen ou la mesure d'un paramètre mesurable. "Négation" marque qu'une entité est notée comme absente, et "Hypothétique" qu'elle constitue une hypothèse, donc n'est pas factuelle.

Les entités « *Vie patient* » permettent d'annoter les comportements du patient (tabagisme, suivi d'un régime...) qui peuvent être favorables ou défavorables à son état. "Entourage" permet de préciser si le patient est accompagné dans la vie de tous les jours ou s'il est isolé. "Autonomie" permet d'indiquer le degré d'autonomie du patient (peut-il marcher sans aide ? est-il capable de descendre des escaliers seul ?).

Les entités « *Localisation* » annotent les lieux où se situe le patient, et ses éventuels déplacements (passage en réanimation, retour à domicile...).

Les entités « *Temporalité* » annotent les données temporelles : heures, dates, âge du patient, durée (d'un traitement, d'un symptôme...), fréquence (d'un traitement, etc.).

4 Expériences

Nous décrivons les procédures d'annotation automatique de nos textes, puis les différentes expériences d'entraînement de modèles supervisés que nous avons réalisées.

4.1 Pré-annotation automatique par lexiques et patrons

Notre système de pré-annotation [2] repose sur des lexiques de termes médicaux en français issus notamment de l'UMLS et de la base de données publique du médicament, que le système cherche à retrouver à l'identique. Il utilise aussi des patrons fondés sur des préfixes et suffixes, des mots-clés et mots déclencheurs qui servent à repérer et à typer les entités des divers types cités plus haut. Il emploie enfin des expressions régulières pour repérer diverses valeurs numériques. Ce système a été appliqué à l'ensemble des dossiers dont nous disposons.

4.2 Entraînement d'un modèle supervisé

Nous suivons l'état de l'art et utilisons un modèle de la famille BERT pré-entraîné sur des textes du domaine médical en français : DrBERT [8], plus précisément le modèle `DrBert-7Gb-Large`. Nous l'utilisons comme encodeur, suivi d'une couche de classification qui étiquette chaque mot selon l'une des classes vues à l'entraînement. Selon le format I-O-B, les mots d'une mention d'entité de type T sont annotés par une classe B-T (*Begin*) pour le premier et I-T (*Inside*) pour les suivants. Les mots qui ne sont dans aucune entité sont annotés O (*Other*). Le texte d'un dossier est traité paragraphe par paragraphe ; dans moins de 1 % des cas, la taille du paragraphe dépasse la taille d'entrée de BERT (512 tokens). Nous n'avons pas effectué de traitement spécial pour gérer ce cas, le comportement par défaut de BERT qui est de tronquer l'entrée à 512 tokens est donc appliqué. Si des entités étaient présentes dans la fin du paragraphe, elles sont donc non vues par le modèle et constituent des faux-négatifs. Il pourrait être utile de découper de telles entrées en paquets de moins de 512 tokens si le cas devenait plus fréquent.

Le classifieur est entraîné pendant un maximum de 50 époques, en ajustant les poids du modèle DrBERT. L'entraînement est automatiquement arrêté après 3 époques sans amélioration des performances. Le taux d'apprentissage débute à 1e-5. La taille de lot est fixée à 8, la dégradation des pondérations à 0,1. Le classifieur est évalué sur un jeu de test annoté manuellement. Nous rapportons les mesures classiques en détection d'entités : précision, rappel, F-mesure des entités avec frontières et types exacts.

4.3 Annotations humaines

Dans les expériences présentées ici, une partie des dossiers pré-annotés automatiquement a ensuite été corrigée par des annotateurs humains :

- 10 textes sont corrigés par une annotatrice experte du domaine ; ancienne assistante de recherche clinique, elle a une connaissance avancée du contenu des dossiers du service de cardiologie ;
- 50 textes sont corrigés par un annotateur non-expert ; doctorant en traitement automatique des langues, il n'a pas de connaissances en médecine. Ces annotations n'ont pas fait l'objet d'un calcul d'accord inter-annotateur.
- Les 60 textes décrits plus haut ont pour limitation l'absence de consensus, c'est-à-dire que ces textes n'ont été annotés que par une seule personne experte, et une seule personne non-experte respectivement.
- Les 10 textes corrigés par l'annotatrice experte ont également été corrigés par l'annotateur non-expert, dans le but de calculer l'accord inter-annotateur entre l'experte et le non-expert sur ces 10 dossiers, afin d'avoir une idée des similitudes entre les deux annotateurs, et de permettre à l'annotateur non-expert d'ajuster sa façon d'annoter. Les 10 textes annotés résultants sont cependant exclus des corpus d'entraînement et de test des expériences menées dans les sections suivantes.

La correction a été effectuée sous le logiciel BRAT [16]. Les annotations expertes seront considérées comme une vérité terrain. Dans toutes les expériences réalisées, le jeu de test sera puisé dans ces textes. Une limitation actuelle est que les textes annotés manuellement par l'experte du domaine ne sont pas des annotations de consensus, car ils sont annotés par une unique experte.

4.4 Entraînement sur des annotations expertes

Notre première expérience a deux objectifs : d'une part évaluer la qualité d'un modèle entraîné uniquement sur peu

d'annotations expertes, d'autre part évaluer la cohérence de l'annotation experte. Nous employons pour cela la technique de validation croisée en dix parties, chaque partie étant composé de 8 dossiers en train, 1 dossier en dev et 1 dossier en val, les dossiers de chaque partie sont ensuite décomposés en phrase. Nous avons choisi de construire les parties de la validation croisée avec des textes entiers plutôt que des ensembles de phrases mélangées pour mieux reproduire la variabilité observée d'un texte à l'autre. Le modèle est chaque fois entraîné sur un maximum de 10 époques, et manuellement arrêté dès qu'aucune amélioration des performances n'est constatée sur 2 époques successives.

Id	Précision	Rappel	F1
1	0,80	0,84	0,82
2	0,55	0,71	0,62
3	0,98	0,98	0,98
4	0,61	0,66	0,63
5	0,74	0,75	0,74
6	0,64	0,66	0,65
7	0,59	0,63	0,59
8	0,57	0,73	0,66
9	0,59	0,58	0,59
10	0,91	0,93	0,92
moy.	0,70 (±0,13)	0,75 (±0,10)	0,72 (±0,12)

TABLE 1 – Validation croisée avec annotations expertes.

La table 1 révèle de fait une grande variabilité des performances du modèle en fonction des données d'entraînement, illustrée par un écart type de 0,12 sur la F-mesure. Celle-ci est en partie attribuable à la faible quantité de dossiers (10) bénéficiant d'une annotation experte. De plus, la diversité des dossiers patients, certains étant plus détaillés en raison de séjours prolongés ou de situations complexes, tandis que d'autres sont plus succincts en raison de pathologies moins sévères, impacte également les résultats obtenus.

4.5 Entraînement sur des pré-annotations uniquement

Nous examinons maintenant la qualité que l'on peut obtenir en entraînant un classifieur exclusivement sur des données pré-annotées automatiquement et en le testant sur les dix dossiers de notre vérité terrain.

Nous testons la contribution de différents sous-ensembles des données pré-annotées en y prélevant des groupes plus ou moins grands, approximativement regroupés par date de séjour des patients. Nous obtenons ces groupes à travers les numéros de séjour attribués aux patients lors de leur admission à l'hôpital. Ces numéros de séjour constituent une indexation chronologique approximative des séjours. La plupart des séjours concernent des patients distincts. Nous créons ainsi deux grands groupes distincts g1 et g2 et y prélevons des sous-groupes g1.1 et g2.1 ainsi que des sous-sous-groupes distincts g1.1.2–5. Aucun des dossiers corrigés manuellement par les experts ou les non-experts n'est présent dans les sous-groupes ainsi créés. La table 2 donne la taille de ces groupes. Elle montre les performances qu'ils permettent d'obtenir, par F-mesure croissante.

Données	Nb. dossiers	P	R	F1
g2	558	0,51	0,57	0,54
g2.1	115	0,52	0,60	0,56
g1.1.4	31	0,55	0,63	0,59
g1.1.3	40	0,60	0,63	0,61
g1.1.2	24	0,61	0,63	0,62
g1.1.5	49	0,60	0,64	0,62
g1.1	350	0,67	0,63	0,65
g1	1508	**0,76**	**0,71**	**0,74**

TABLE 2 – Performances avec entraînement uniquement sur la pré-annotation.

On voit que les groupes g1 ont une meilleure performance que les groupes g2. Par ailleurs, un plus grand nombre de dossiers du groupe g1 améliore la performance alors que c'est l'inverse au sein du groupe g2. Il semble donc que les dossiers du groupe g1 ont une pré-annotation qui est plus en accord avec les annotations attendues pour les dossiers du test. Par ailleurs, un grand nombre de g1 améliore davantage la précision et la fait passer au-dessus du rappel, ce qui reste à expliquer. Enfin, la précision et la F-mesure obtenues pour le groupe g1 (P=0,76, F1=0,74) dépassent celles obtenues par le modèle entraîné en validation croisée sur les dossiers à correction experte (P=0,70, F1=0,72). Cependant, l'expérience g1 n'a été réalisée qu'une fois, sa répétition avec des amorces différentes pourrait donc être cause de variance.

4.6 Entraînement sur des annotations non-expertes et des pré-annotations

Dans l'expérience précédente, nous avons mesuré les performances du modèle sur les dix dossiers annotés par l'experte du domaine médical. Les résultats obtenus ont permis d'évaluer l'efficacité initiale de notre approche. Nous examinons ici si l'ajout de données faisant l'objet de corrections non-expertes aux données pré-annotées peut mener à l'entraînement d'un modèle possédant de meilleures performances. Pour cette nouvelle expérience, nous appliquons la même méthode utilisée précédemment, mais cette fois-ci sur un ensemble de cinquante dossiers annotés par un non-expert. Le test se fait toujours sur les dix dossiers de notre vérité terrain.

La table 3 montre la performance du modèle entraîné sur les annotations non-expertes (50 textes), puis celles lorsqu'on y ajoute les mêmes groupes et sous-groupes. On voit que 50 dossiers avec correction non-experte (table 3, Non-expert) font mieux que plusieurs centaines de dossiers pré-annotés sans correction (table 2, toutes lignes sauf g1). Néanmoins, l'ajout de ces 50 dossiers aux annotations sans correction améliore toujours les performances du modèle entraîné sur ces données (colonne ΔF1 de la table 3). Cette amélioration est relativement stable, de l'ordre de 4 à 5 points de F-mesure. Comme précédemment, un grand nombre de dos-

Données	Nb. dossiers	P	R	F1	ΔF1
Non-expert	50	0,64	0,70	0,67	
N.E. + g2	+558	0,56	0,63	0,59	+0,05
N.E. + g1.1.4	+31	0,60	0,69	0,64	+0,05
N.E. + g2.1	+115	0,61	0,69	0,65	+0,09
N.E. + g1.1.3	+40	0,62	0,69	0,65	+0,04
N.E. + g1.1.5	+49	0,62	0,69	0,65	+0,03
N.E. + g1.1.2	+24	0,63	0,69	0,66	+0,04
N.E. + g1.1	+350	0,71	0,67	0,69	+0,04
N.E. + g1	+1508	**0,81**	**0,76**	**0,79**	+0,05

TABLE 3 – Performances avec entraînement sur annotations non-expertes et pré-annotations. ΔF1 : différence de F1 par rapport aux expériences avec pré-annotation uniquement.

siers du groupe g2 mène à des performances inférieures, et un grand nombre de dossiers du groupe g1 augmente les performances, en faisant passer la précision au-dessus du rappel. Les performances du groupe N.E.+g1 dépassent sensiblement celles de la validation croisée sur les dossiers à correction experte.

4.7 Évaluation par classe de la performance du meilleur modèle

Évaluer la performance d'un modèle pour chaque classe du schéma d'annotation nous permet d'apprécier plus précisément les forces et faiblesses de ce modèle y compris pour les classes moins représentées. Nous effectuons ce calcul pour le meilleur modèle identifié dans la table 3 (N.E. + g1). La table 4 montre les résultats.

Le modèle se distingue particulièrement dans la classification des entités de la classe très fréquente "Traitement", où précision et rappel surpassent les performances moyennes du modèle. Cela peut s'expliquer par le fait que les entités de cette classe présentent fréquemment des motifs distinctifs dans nos données, tels que des noms de médicaments en majuscules, des listes de médicaments en tête de texte, ou des suffixes spécifiques comme -ectomie.

Les entités de type "Valeur" sont plus délicates à identifier pour le modèle. Cela pourrait être lié à la complexité à les distinguer d'autres types de valeurs numériques, comme des dates ou des concentrations, par des règles dans la pré-annotation puis par le classifieur entraîné.

Les entités relativement fréquentes suivantes, comme les pathologies, dates, signes/symptômes, concentrations, doses, paramètres mesurables, etc., ont des performances qui se rapprochent des valeurs moyennes du modèle.

Les performances sont plus faibles pour des classes telles qu'Évolution et Examen. Cela est vraisemblablement dû à une représentation insuffisante de ces classes, qui demeurent relativement rares avec seulement une dizaine d'occurrences chacune dans le test.

Les classes les moins fréquentes tendent à afficher des performances extrêmes (très hautes ou très basses) en raison de leur faible occurrence. Néanmoins, leur influence sur les performances globales demeure de ce fait limitée. De plus, les classes peu fréquentes et moins bien détectées ne font pas partie des catégories essentielles susceptibles de modifier radicalement le sens d'une phrase, réduisant ainsi le risque d'un impact négatif sur les performances globales.

Classe	Nb	P	R	F1
Micro-moyenne	690	0,81	0,76	0,79
Traitement	155	**0,90**	**0,82**	**0,85**
Valeur	59	0,67	0,69	0,68
Pathologie	58	0,79	**0,76**	0,77
Date	54	**0,87**	0,70	**0,79**
Signe/Sympt.	53	0,80	0,74	0,77
Concent.	46	0,78	**0,82**	**0,80**
Dose	43	0,77	**0,92**	**0,85**
Anatomie	42	**0,87**	0,58	0,72
Param. mes.	38	0,77	**0,79**	0,78
Fréquence	29	0,75	**0,83**	**0,79**
Négation	20	**1,00**	0,73	**0,87**
Mode	15	**0,86**	**0,86**	**0,86**
Lieu	15	0,70	**0,81**	0,75
Hypothèse	13	**1,00**	0,58	**0,79**
Évolution	12	0,50	0,50	0,50
Examen	10	0,59	0,63	0,61
Entourage	6	**0,87**	**1,00**	**0,94**
Heure	5	**1,00**	**0,80**	**0,90**
Chgt. lieu	5	0,67	**0,80**	0,74
Comport.	4	0,67	**1,00**	**0,84**
Âge	3	**1,00**	**1,00**	**1,00**
Durée	3	0,67	0,33	0,50
Autonomie	2	0,68	**1,00**	**0,80**

TABLE 4 – Performances du meilleur modèle (N.E. + g1) par classe, par ordre décroissant du nombre d'entités. Les valeurs supérieures ou égales à la moyenne sont en gras.

5 Résultats et discussion

Nous avons entraîné des modèles de détection d'entités avec plusieurs sortes de données annotées et les avons évalués sur des textes vus par une annotatrice experte.

Avec dix dossiers avec corrections expertes, la performance du modèle reste limitée (validation croisée sur 10 dossiers, F1=0,72).

De très nombreuses pré-annotations non corrigées (pré-annotation uniquement, groupe g1 sur 1508 dossiers), donc peu coûteuses, peuvent faire mieux (F1=0,74).

Des annotations non-expertes sur 50 dossiers n'atteignent pas ce résultat. En revanche, l'ajout de ces annotations non-expertes aux très nombreuses pré-annotations (de nouveau, groupe g1) accroît leur performance (F1=0,79).

L'analyse par classe relève d'une part des entités plutôt bien reconnues, notamment autour des posologies (traitement, concentration, dose, fréquence, mode), et d'autre part des entités qui bien que non rares, ont une reconnaissance moins bonne (valeur, évolution, examen).

Le peu de données avec correction experte limite la taille du test et notre capacité d'interprétation des résultats pour les classes moins fréquentes. C'est une limite des expériences réalisées. Celle-ci se réduira lorsque davantage de dossiers auront été annotés par l'annotatrice experte. L'absence de double-annotation experte est une autre limite actuelle.

6 Conclusion

Dans cet article, nous avons exploré différentes stratégies d'annotation pour améliorer les performances d'un modèle de reconnaissance d'entités nommées dans le domaine médical en présence de très peu d'annotations expertes. Nos résultats indiquent que dans ce contexte, l'utilisation de données pré-annotées non-corrigées, mais peu coûteuses et disponibles en grande quantité, peut mener à l'entraînement d'un classifieur plus performant, et que des annotations non-expertes en quantité modérée peuvent les compléter. Toutefois, la nature des données utilisées est un facteur important à considérer, et son étude est à approfondir. De plus, la quantité limitée d'annotateurs (un non-expert, un expert) à notre disposition limite la fiabilité des annotations, y compris pour le jeu de test : son amélioration devra être visée en priorité.

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

[1] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In Anna Rumshisky, Kirk Roberts, Steven Bethard, and Tristan Naumann, editors, *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[2] Virgile Barthet, Marie-José Aroulanda, Laura Monceaux-Cachard, Christine Jacquin, Cyril Grouin, Johann Gutton, Guillaume Hocquet, Pascal De Groote, Michel Komajda, Emmanuel Morin, and Pierre Zweigenbaum. 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. In Florian Boudin, Béatrice Daille, Richard Dufour, Oumaima El, Maël Houbre, Léane Jourdan, and Nihel Kooli, editors, *Actes de CORIA-TALN 2023. Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023*, pages 1–7, Paris, France, 6 2023. ATALA.

[3] Louise Deléger, Leonardo Campillos, Anne-Laure Ligozat, and Aurélie Névéol. Design of an extensive information representation scheme for clinical narratives. *J Biomed Semantics*, 8(1) :37, Sep 11 2017.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT : Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[5] Félix Gaschi, Xavier Fontaine, Parisa Rastin, and Yannick Toussaint. Multilingual clinical NER : Translation or cross-lingual transfer ? In Tristan Naumann, Asma Ben Abacha, Steven Bethard, Kirk Roberts, and Anna Rumshisky, editors, *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 289–311, Toronto, Canada, July 2023. Association for Computational Linguistics.

[6] Ralph Grishman and Beth Sundheim. Message Understanding Conference- 6 : A brief history. In *COLING 1996 Volume 1 : The 16th International Conference on Computational Linguistics*, 1996.

[7] Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. Thinking about GPT-3 in-context learning for biomedical IE ? think again. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics : EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[8] Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. DrBERT : A robust pretrained model in French for biomedical and clinical domains. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 16207–16221, Toronto, Canada, July 2023. Association for Computational Linguistics.

[9] Wenhui Liao and Sriharsha Veeramachaneni. A simple semi-supervised algorithm for named entity recognition. In Qin Iris Wang, Kevin Duh, and Dekang Lin, editors, *Proceedings of the NAACL HLT 2009 Workshop on Semi-supervised Learning for Natural Language Processing*, pages 58–65, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[10] Todd Lingren, Louise Deleger, Katalin Molnar, Haijun Zhai, Jareen Meinzen-Derr, Megan Kaiser, Laura Stoutenborough, Qi Li, and Imre Solti. 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 Health-*

care Informatics, Imaging and Systems Biology, pages 108–108, 2012.

[11] Pierre Lison, Jeremy Barnes, and Aliaksandr Hubin. skweak : Weak supervision made easy for NLP. In Heng Ji, Jong C. Park, and Rui Xia, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing : System Demonstrations*, pages 337–346, Online, August 2021. Association for Computational Linguistics.

[12] Danielle L Mowery, Pamela Jordan, Janyce Wiebe, Henk Harkema, John Dowling, and Wendy W Chapman. Semantic annotation of clinical events for generating a problem list. *AMIA Annu Symp Proc*, 2013 :1032–1041, 2013.

[13] Marco Naguib, Xavier Tannier, and Aurélie Névéol. Few shot clinical entity recognition in three languages : Masked language models outperform LLM prompting. *CoRR*, abs/2402.12801, 2024.

[14] Lisa Raithel, Philippe Thomas, Roland Roller, Oliver Sapina, Sebastian Möller, and Pierre Zweigenbaum. Cross-lingual approaches for the detection of adverse drug reactions in German from a patient's perspective. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3637–3649, Marseille, France, June 2022. European Language Resources Association.

[15] Emiko Shinohara, Daisaku Shibata, and Yoshimasa Kawazoe. Development of comprehensive annotation criteria for patients' states from clinical texts. *J Biomed Inform*, 134 :104200, Oct 2022.

[16] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. brat : a web-based tool for NLP-assisted text annotation. In Frédérique Segond, editor, *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France, April 2012. Association for Computational Linguistics.