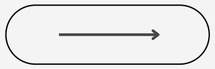
Équilibrer qualité et quantité : comparaison de stratégies d'annotation pour la reconnaissance d'entités nommées en cardiologie



GROUPE HOSPITALIER PARIS SAINT-JOSEPH

CHU LILLE

LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

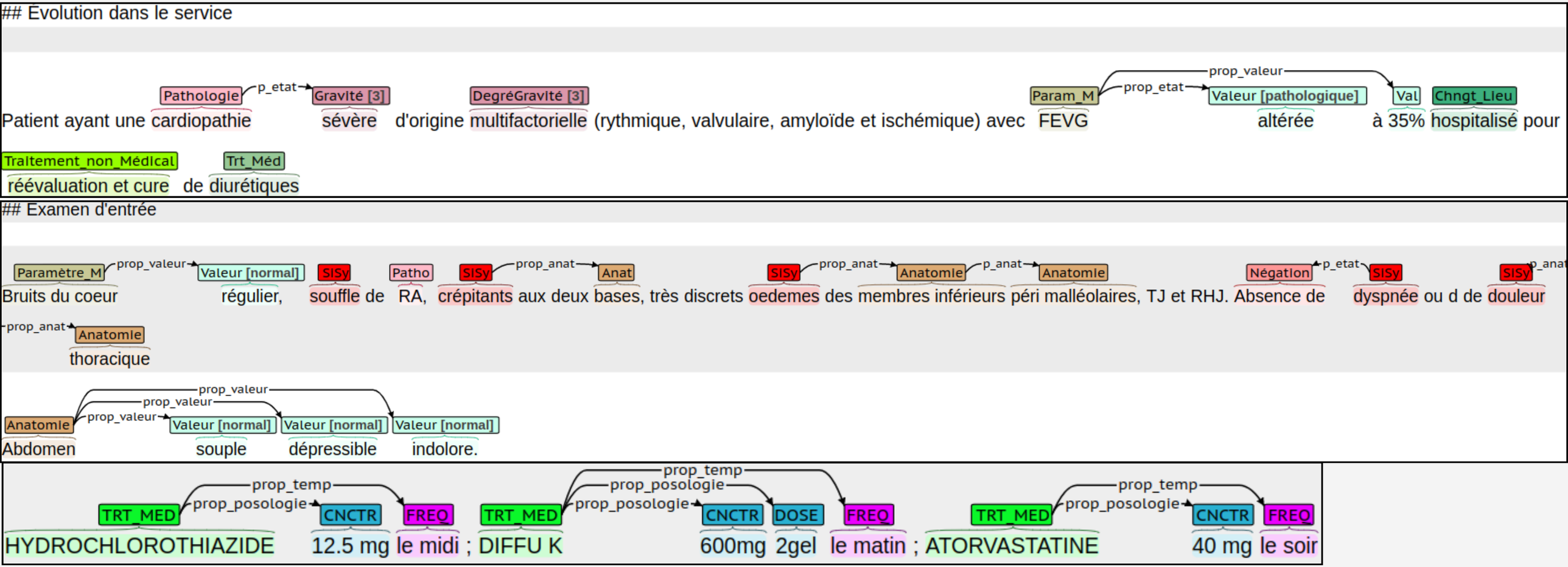
LS2N

cnrs dépasser les frontières

anr

université PARIS-SACLAY

Détection d'entités médicales



Détection d'entités avec peu d'annotations

- Il est difficile d'obtenir des annotations de qualité
  - L'annotation manuelle est un processus laborieux
  - L'annotation requiert une expertise rare

- Annotation en deux étapes
  - Pré-annotation automatique (Barthet et al., 2023)
  - Correction humaine

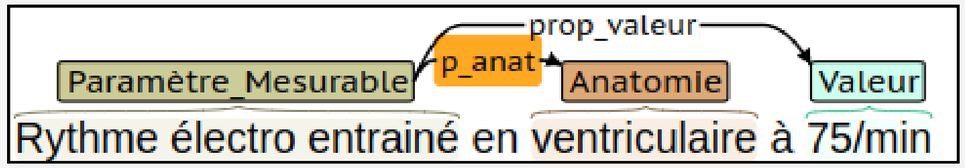
- Entraîner un système de reconnaissance d'entités nommées
  - Quel est le minimum d'annotations corrigées pour y arriver ?
  - La pré-annotation sans correction peut-elle être utile ?

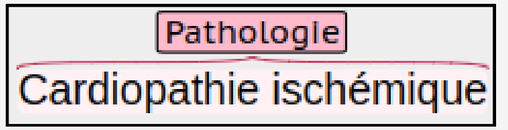
Contexte : Projet predhic

- Projet ANR : traitement automatique des langues pour la cardiologie
- Collaboration avec les services de cardiologie de 2 établissements
  - Le Groupe Hospitalier Paris Saint-Joseph
  - Le CHU de Lille
- Accès aux notes cliniques de cardiologie des patients consentants
- But final: prédiction de risques de réhospitalisation ou décès des patients atteints d'insuffisance cardiaque, par apprentissage supervisé
- Étape préalable: annotation des données pour entraîner un modèle

Données

- Notes cliniques
  - Phrases en langue naturelle
  - Vocabulaire très spécialisé, abréviations fréquentes





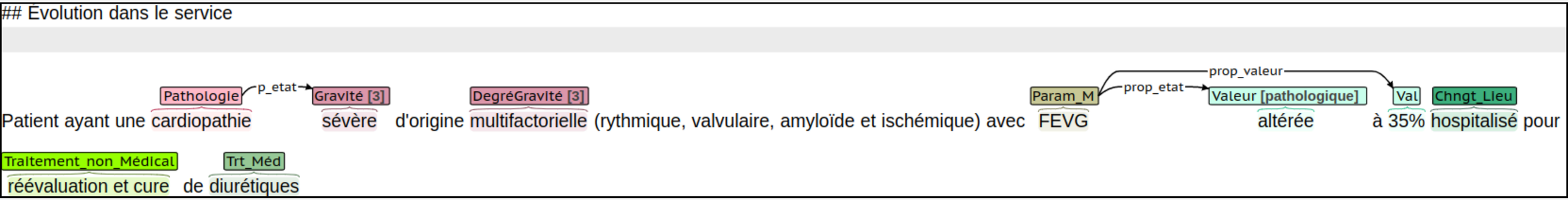
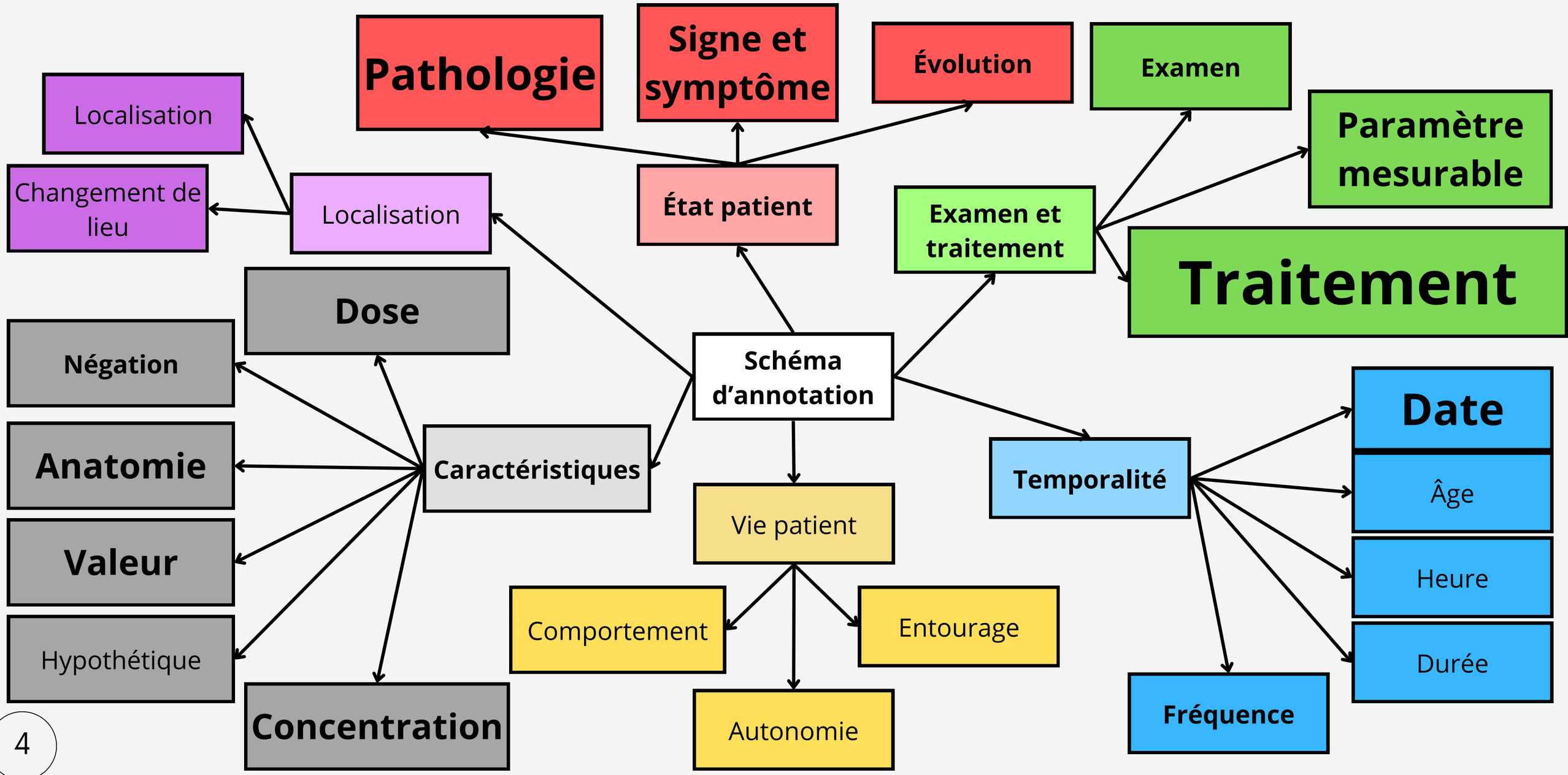


Schéma d'annotation

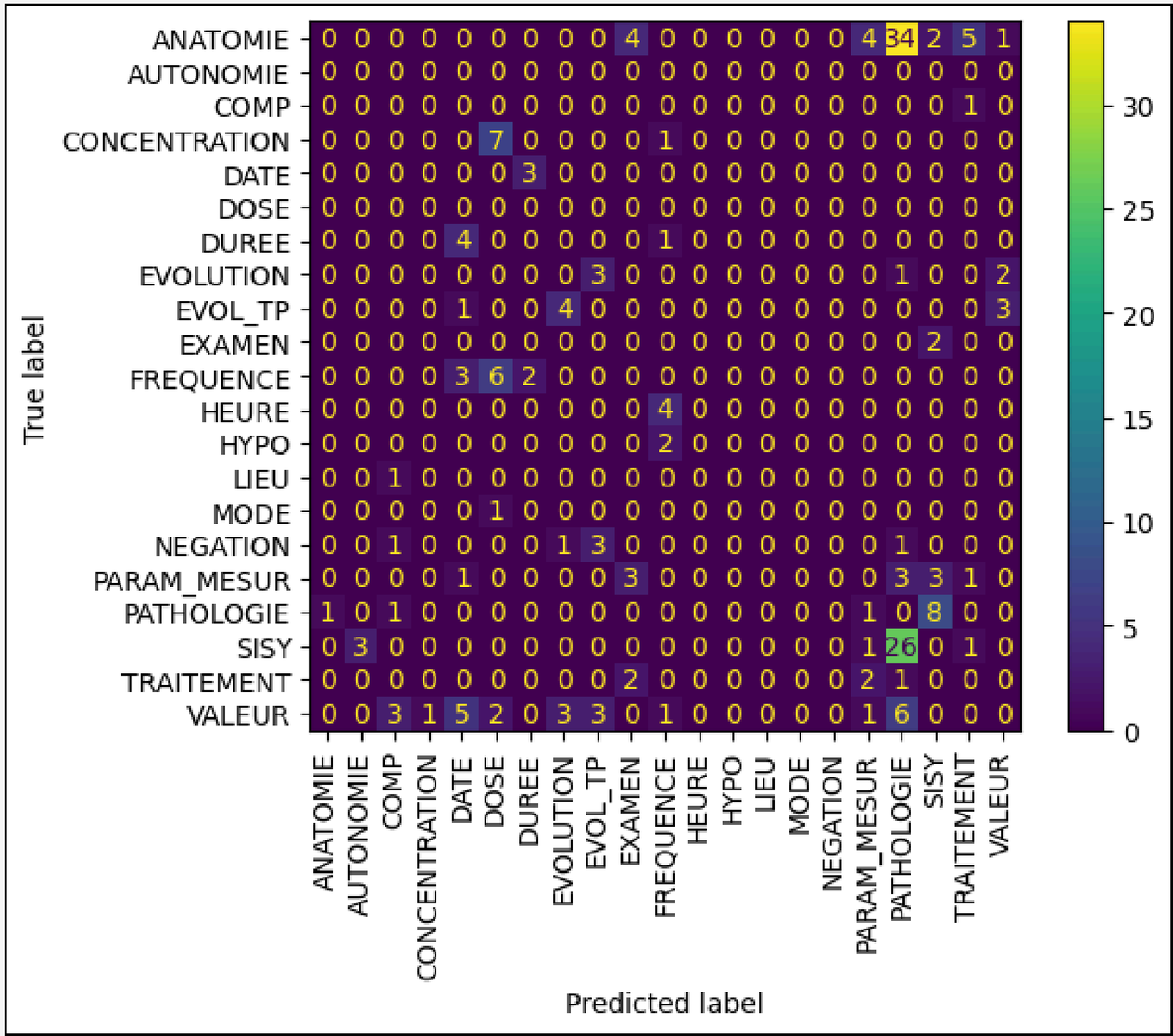


Trois qualités d'annotation

Méthode d'annotation	Annotateur	Qualité	Quantité de textes
Pré-annotation	Programme automatisé basé sur des règles, dictionnaires, regexp ...	moyenne	> 1000
Correction non-experte	Doctorant en TAL qui corrige la pré-annotation	mieux que la pré-annotation, moins bien que l'annotation experte	50
Correction experte	Experte en cardiologie qui corrige la pré-annotation	vérité terrain	10

Non-expert vs experte : matrice de confusion

- Mesure de la qualité de la correction non-experte par rapport à l'experte (considérée comme référence)
- Accord inter-annotateur
- Permet de rectifier la manière d'annoter non-experte
- Permet d'identifier des défauts du schéma d'annotation



Entraînement d'un classifieur fondé sur BERT

- Apprentissage par transfert

- Modèle BERT pré-entraîné sur des textes médicaux en français
  - DrBert-7Gb-Large
  - Ajout d'une tête de classification entraînée sur nos données
  - Annotations au format BIO (Begin, Inside, Other)

- Évaluation sur les textes de la correction experte : précision, rappel et F-mesure.

Entraînement sur la correction experte seule

Jeu de données	Nombre de textes
Train	8
Dev	1
Test	1

- Validation croisée pour compenser la faible quantité de textes
- Standard auquel se comparer ensuite

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)

Entraînement sur les données pré-annotées

Les numéros de séjour sont une suite de 9 chiffres (exemple: 919237528)

SOUS-GROUPES DE DONNÉES PRÉ-ANNOTÉES

Basés sur les numéros de séjour:

- Sous-groupe des numéros débutants par “11” = groupe g1
- Sous-groupe des numéros débutants par “91” = groupe g2
- Etc.
- Numéros de séjour ~= Dates de séjour

SÉPARATION EN SOUS-SOUS-GROUPES

Parmi le sous-groupe débutant par “11”:

- Numéros débutant par “1160” = groupe g1.1
- Numéros débutant par “1162” = groupe g1.2
- Numéros débutant par “11603” = groupe g1.1.3
- Numéros débutant par “11604” = groupe g1.1.4
- Etc.

Entraînement sur les données pré-annotées

Jeu de données	Nombre de textes
Train: Pré-annotation	Variable
Test: Correction experte	10

- Sous-groupes tous de même qualité ?

- → Performances très variables mais g1 > g2
- → Performances globalement insuffisantes
- → Performances > exp. 1 pour g1

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**

Entraînement sur non-expert + pré-annotation

Jeu de données	Nombre de textes
Train: Correction non-experte + Pré-annotation	50 + [24 .. 1508]
Test: Correction experte	10

- → Toujours g1 > g2
- → La combinaison améliore les performances du classifieur

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

Performances par classe

Jeu de données	Nombre de textes
Train: Correction non-experte + Sous-groupe g1	1558 (50 + 1508)
Test: Correction experte	10

- → Performances inégales selon les classes
- → Plus de difficultés sur les valeurs
- → Très bonnes performances sur la classe la plus fréquente (Traitement)

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**

CLASSES < 20 OCCURRENCES NON AFFICHÉES

Discussion

- Performances limitées si correction experte seule

- Les pré-annotations en grande quantité peuvent faire mieux que peu de correction experte, tout en étant moins coûteuses

- L'ajout d'annotations non-expertes aux pré-annotations améliore les performances

- Des entités mieux reconnues
  - Traitement : syntaxe régulière et noms faciles à reconnaître
  - Date : facile à reconnaître
- Des entités moins bien reconnues
  - Valeur : confusion avec d'autres entités (dose, concentration)

Discussion

- Les classes moins fréquentes (voir Table 4 de l'article) tendent à afficher des performances plus extrêmes (très hautes ou très basses) en raison de leur faible occurrence

- Le peu de corrections expertes disponibles limite la capacité d'interprétation des résultats pour les classes moins fréquentes

- L'absence de double-annotation experte limite le degré de fiabilité de la correction experte

MERCI

PRÉSENTÉ PAR

BARTHET Virgile,
doctorant au LISN

GROUPE HOSPITALIER PARIS SAINT-JOSEPH

CHU LILLE

LISN
LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

LS2N

cnrs dépasser les frontières

anr

université PARIS-SACLAY