

Enrichir une ontologie médicale avec des connaissances de résumés de sortie de l'hôpital

S. Després, C. Duclos, C. Le Duc et P. Vaillant
Journée de Santé et IA, 2023



Motivations

- Exploitation d'une ontologie médicale peuplée (ABox) avec des services de raisonnement et requêtes (SPARQL, conjonctives)
- Sources de textes médicaux riches et accessibles (EHR, DS) avec des connaissances factuelles
- Outils TAL (classique et apprentissage profond) : cTakes, SpaCy
- Outils ontologiques : raisonneurs

Exemple 1

Une ontologie \mathcal{O} sans individu

PatientSufferFromPneumonia $\sqsubseteq \exists$ sufferFrom.Pneumonia

PatientMedicatedWithAmox $\sqsubseteq \exists$ medicatedWith.Amoxicillin

PatientAllergicToAmox $\sqsubseteq \forall$ medicatedWith. \neg Amoxicillin

On peuple \mathcal{O} en ajoutant les assertions suivantes :

PatientSufferFromPneumonia(*Sally*)

PatientMedicatedWithAmox(*Sally*)

PatientAllergicToAmox(*Sally*)

L'ontologie peuplée n'est pas cohérente

Exemple 2

Une ontologie peuplée :

...

Aspirin \sqsubseteq Anti-inflammatoryAgent

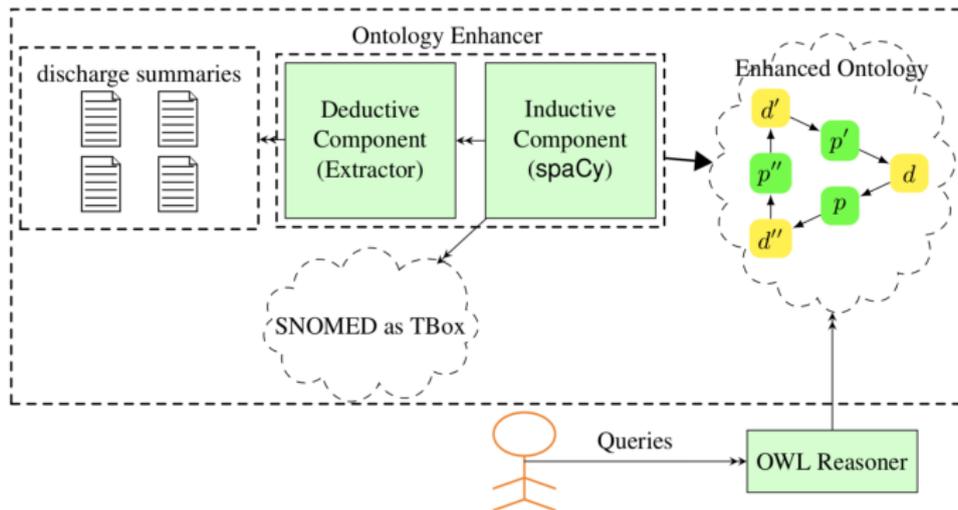
Aspirin \sqsubseteq Salicylate

CardiacPatient \sqsubseteq \exists medicatedWith.Aspirin

CardiacPatient(*Trude*)

Chercher tous les médicaments/substances prescrits à Trude

L'approche : architecture



L'approche : étape 1

```
RECORD #106886
Admission Date: 4/8/1994
Discharge Date: 10/20/1994
HISTORY OF PRESENT ILLNESS: Patient is a 73 year old male who immigrated
from ...
ALLERGIES: Penicillin caused a rash.
PRINCIPAL DIAGNOSIS: STATUS POST STAPHYLOCOCCUS ENDOCARDITIS
DISCHARGE DIAGNOSIS: Diabetes, INFECTION/RULE OUT ENDOCARDITIS
...
DISCHARGE MEDICATIONS: Tylenol 650 mg p.o. q.4h. p.r.n. headache ,
Ventolin two puffs inhaled b.i.d. , Beclovent two puffs inhaled b.i.d. ,
20 units subcutaneous q.a.m.; Penicillin B 250 mg p.o. q.i.d. ...
```

Utiliser les outils classiques basés sur les expressions régulières :

- Découper chaque DS en sections (ALLERGIES, PRINCIPAL DIAGNOSIS, DISCHARGE DIAGNOSIS, DISCHARGE MEDICATIONS)
- Réduire la taille de texte pour le traitement par d'autres outils (spaCy)

L'approche : étape 2

```
RECORD #106886
Admission Date: 4/8/1994
Discharge Date: 10/20/1994
HISTORY OF PRESENT ILLNESS: Patient is a 73 year old male who immigrated
from ...
ALLERGIES: Penicillin caused a rash.
PRINCIPAL DIAGNOSIS: STATUS POST STAPHYLOCOCCUS ENDOCARDITIS
DISCHARGE DIAGNOSIS: Diabetes, INFECTION/RULE OUT ENDOCARDITIS
...
DISCHARGE MEDICATIONS: Tylenol 650 mg p.o. q.4h. p.r.n. headache ,
Ventolin two puffs inhaled b.i.d. , Beclovent two puffs inhaled b.i.d. ,
20 units subcutaneous q.a.m.; Penicillin B 250 mg p.o. q.i.d. ...
```

Détection de relations : assertions et axiomes

Patient(#106886)

{#106886} \sqsubseteq \exists sufferFrom. Diabetes

{#106886} \sqsubseteq \exists medicatedWith. Penicillin

{#106886} \sqsubseteq \exists allergicTo. Penicillin

L'approche : étape 3 avec cTakes

cTakes (Clinical Text Analysis and Knowledge Extraction System)

- Détecter les phrases (ponctuation)
- “Tokeniser” une phrase en groupes de mots liés (dates, nom de personnes, fractions, mesures, etc.)
- Normaliser des tokens (les variants d'orthographe)
- Mettre en correspondance entre les tokens trouvés et ceux définis dans les terminologies standardisées (UMLS)
- Détecter les tokens de négation en se basant sur des patrons syntaxiques (RULE OUT, DENY, etc.)

L'approche : étape 4 avec spaCy

spaCy fondé sur une méthode statistique (apprentissage profond)

- Tokenisation et normalisation (dictionnaires)
- *word vectors* : représentation du contexte d'un terme (textes d'entraînement)
- *feature function* : prédiction des correspondances entre les termes (*word vectors*) et les entités
- Sélection (*représentativité*) et annotation manuelle (textes d'entraînement)
- spaCy reconnaît une entité dans un nouveau texte (*feature function*)

● Analyse des DS

Dataset of 696 DS	Nb of axioms	Nb of assertions
DIAGNOSIS ON ADM/DISC	900	2,214
MEDICATIONS ON ADM/DISC	609	4,512
ALLERGY	372	878

	Precision	Recall
Diagnostics Extraction	0.72	0.61
Medication Extraction	0.80	0.59

La faible mesure de rappel dû à des sections pas bien délimitées, et donc tronquées.

● Analyse des DS

Dataset of 696 DS	Nb of axioms	Nb of assertions
DIAGNOSIS ON ADM/DISC	900	2,214
MEDICATIONS ON ADM/DISC	609	4,512
ALLERGY	372	878

	Precision	Recall
Diagnostics Extraction	0.72	0.61
Medication Extraction	0.80	0.59

La faible mesure de rappel dû à des sections pas bien délimitées, et donc tronquées.

- Un fragment de SNOMED (patient/medicament/maladie) peuplée consultable et interrogeable en ligne

Conclusion et discussion

- Conclusion
 - Une approche pour le peuplement d'une ontologie médicale
 - Un fragment de SNOMED-CT (patient/médicament/maladie) peuplé avec 700 CRH
 - Ontologie peuplée accessible et interrogeable
- Discussion
 - Augmentation de l'expressivité de SMOMED ?
 - Application de la méthode à une autre ontologie ?
 - Autres sources de textes et autres outils du TAL ?
 - Fondement de l'approche : utilisation d'un outil statistique dans le cadre symbolique ?