Atelier Inaugural

Data science, Health data & AI
TAL en Santé

Xavier Tannier et Perceval Wajsbürt

1. Traitement automatique des langues et données de santé

Xavier Tannier

1. Structuration de données textuelles

   Cadre générique d'un projet nécessitant l'extraction d'information dans les comptes-rendus textuels.

2. Focus sur la normalisation des concepts médicaux dans les textes

   Un projet de recherche en traitement automatique des langues

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft*

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Extraction d'information à partir des comptes-rendus textuels

- **Objectif** : structurer des informations présentes à l'origine uniquement dans des parties non structurées des comptes-rendus cliniques (texte)

- … se décline en objectifs en termes de
  - **Collaboration** : Faciliter le dialogue entre spécialistes du domaine concerné, experts en traitement automatique des langues, experts de l'entrepôt de données
  - **Coordination** : Mutualiser les efforts entre les projets
  - **Méthodes** : Réduire l'effort en termes d'annotation et d'ingénierie
  - **Évaluation** : Identifier la fiabilité des informations extraites
  - **Déploiement** : Rendre possible la mise à disposition sur l'entrepôt de données de santé

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft*

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Différentes tâches

1. Extraction de concepts
2. Normalisation de concepts
3. Extraction de relations
4. Classification de documents
5. Classification de patients

- Cette classification des tâches n'est pas un ordre de difficulté de ces tâches, mais en général les tâches d'en bas (moins ancrées dans le texte) impliquent la résolution de tâches d'en haut.
- Chaque étape doit être évaluée.

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft*

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Différentes tâches

Extraction et caractérisation de concepts

- Extraction des entités nommées

MAMMOGRAPHIE :

On décèle la présence de foyers de micro-calcifications (LÉSION) (LOCALISATION) (LOCALISATION)

dans le sein droit (LOCALISATION) dans le rayon de 8h, (LOCALISATION) et dans

le sein gauche dans le rayon de 7h.

CONCLUSION :

L'examen est donc reclassé ACR 4 (SCORE) de chaque côté. (LOCALISATION)

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Différentes tâches

Extraction et caractérisation de concepts

- Caractérisation des entités nommées

CONCLUSION_COVID [diagnostic:probable]

Lésion pulmonaire fortement évocatrice du Covid

PROBLÈME [factualité:neg]

Patient sans signe clinique évident de traumatisme crânien

PROCÉDURE [factualité:antécédent]

Patiente avec antécédent de chirurchie bariatrique

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft*

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Différentes tâches

Extraction et caractérisation de concepts: comment ?

1. Approche terminologique

MALADIE VASCULAIRE

- *claudication intermittente*
- *AOMI*
- *artériopathie des membres*
- *artérite des membres*
- *ulcère artériel*
- *ulcère veineux*
- *insuffisance artérielle*
- *gangrène*
- *ischémie aigue du membre*
- *ischémie aigue périphérique*
- *anévrysme de l'aorte*
- *anévrysme aortique*

Recherche (approximative) des termes dans les documents

MALADIE VASCULAIRE

Patient atteint d'ulcères artériels des membres inférieurs, suivis à St Joseph (Dr Wyliana) avec greffe cutanée en octobre 2015

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft*

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

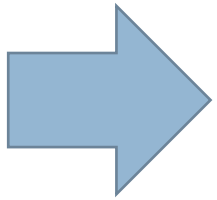
Différentes tâches

Extraction et caractérisation de concepts: comment ?

1. Approche terminologique
2. Ajout de règles

RÉSISTANCE BACTÉRIE

- *<BACTÉRIE> + .* sensible*
- *<BACTÉRIE> + .* résistance*
- *<BACTÉRIE> + sensible à la méthicilline*
- *…*



9 hémocultures positives le 26/6/15 à staphylocoque aureus méticilline sensible.

RÉSISTANCE BACTÉRIE

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft*

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Différentes tâches

Extraction et caractérisation de concepts: comment ?

1. Approche terminologique
2. Ajout de règles
3. Si besoin, apprentissage supervisé (extraction d'entités nommées)
    - Annotation manuelle de documents
    - Entraînement d'un modèle de reconnaissance de séquences de mots
    - Application de ce modèle à des textes nouveaux

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft*

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Différentes tâches

Normalisation de concepts

[C0018674] (Craniocerebral Trauma)

Patient sans signe clinique évident de traumatisme crânien

[C1456587] (Bariatric Surgery)

Patiente avec antécédent de chirurchie bariatrique

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft*

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Différentes tâches

Normalisation de concepts: comment ?

1. Approche terminologique simple
2. Traduction + approche terminologique
3. Apprentissage:
    - Annotation manuelle de documents (optionnel)
    - Entraînement d'un modèle
    - Application de ce modèle à des textes nouveaux

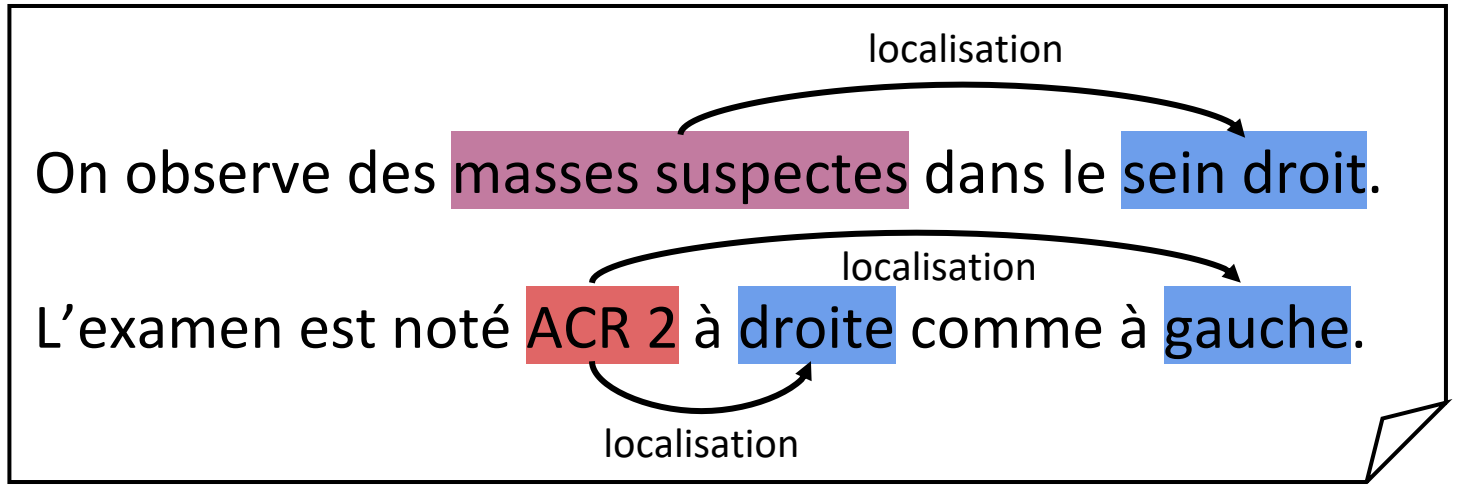
ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft*

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Différentes tâches

Extraction de relations



ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft*

scai
SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Différentes tâches

Extraction de relations: comment ?

1. Règles
2. Apprentissage supervisé
3. Apprentissage distant

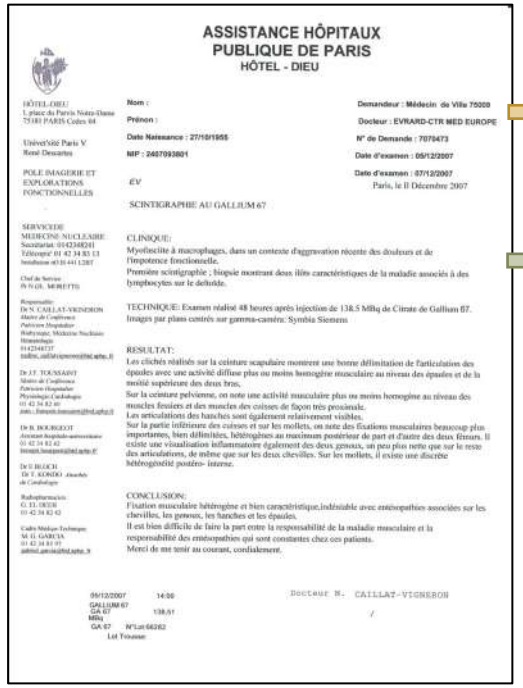
ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft*

scai
SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Différentes tâches

Phénotypage simple: classification de document



ASSISTANCE HÔPITAUX
PUBLIQUE DE PARIS
HÔTEL - DIEU

SCINTIGRAPHIE AU GALLIUM 67

CLINIQUE:
Myofasciite à macrophages, dans un contexte d'aggravation récente des douleurs et de l'impotence fonctionnelle.
Première scintigraphie ; biopsie montrant deux îlots caractéristiques de la maladie associés à des lymphocytes sur le deltoïde.

TECHNIQUE: Examen réalisé 48 heures après injection de 138.5 MBq de Citrate de Gallium 67.
Images par plans centrés sur gamma-caméra: Symbia Siemens

RESULTAT:
Les clichés réalisés sur la ceinture scapulaire montrent une bonne délimitation de l'articulation des épaules avec une activité diffuse plus ou moins homogène musculaire au niveau des épaules et de la moitié supérieure des deux bras.
Sur la ceinture pelvienne, on note une activité musculaire plus ou moins homogène au niveau des muscles fessiers et des muscles des cuisses de façon très proximale.
Les articulations des hanches sont également relativement visibles.
Sur la partie inférieure des cuisses et sur les mollets, on note des fixations musculaires beaucoup plus importantes, bien délimitées, hétérogènes au maximum postérieur de part et d'autre des deux fémurs. Il existe une visualisation inflammatoire également des deux genoux, un peu plus nette que sur le reste des articulations, de même que sur les deux chevilles. Sur les mollets, il existe une discrète hétérogénéité postéro- interne.

CONCLUSION:
Fixation musculaire hétérogène et bien caractéristique,indéniable avec entésopathies associées sur les chevilles, les genoux, les hanches et les épaules.
Il est bien difficile de faire la part entre la responsabilité de la maladie musculaire et la responsabilité des entésopathies qui sont constantes chez ces patients.
Merci de me tenir au courant, cordialement.

Docteur N. CAILLAT-VIGNERON

Motif d'hospitalisation = fracture de faible cinétique ?

Traitement de sortie ?

En lien avec l'extraction et la caractérisation de concepts, mais implique une décision/prédiction au niveau du document.

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

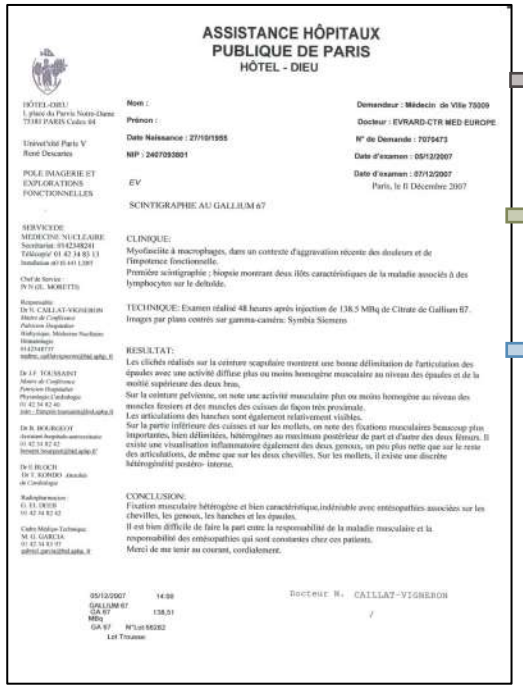
datacraft*

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Différentes tâches

Phénotypage simple: classification de patient

5. Classification de patient / phénotypage simple



ASSISTANCE HÔPITAUX
PUBLIQUE DE PARIS
HÔTEL - DIEU

SCINTIGRAPHIE AU GALLIUM 67

CLINIQUE:
Myofasciite à macrophages, dans un contexte d'aggravation récente des douleurs et de l'impotence fonctionnelle.
Première scintigraphie ; biopsie montrant deux îlôts caractéristiques de la maladie associés à des lymphocytes sur le deltoïde.

TECHNIQUE: Examen réalisé 48 heures après injection de 138,5 MBq de Citrate de Gallium 67.
Images par plans centrés sur gamma-caméra: Symbia Siemens

RESULTAT:
Les clichés réalisés sur la ceinture scapulaire montrent une bonne délimitation de l'articulation des épaules avec une activité diffuse plus ou moins homogène musculaire au niveau des épaules et de la moitié supérieure des deux bras.
Sur la ceinture pelvienne, on note une activité musculaire plus ou moins homogène au niveau des muscles fessiers et des muscles des cuisses de façon très proximale.
Les articulations des hanches sont également relativement visibles.
Sur la partie inférieure des cuisses et sur les mollets, on note des fixations musculaires beaucoup plus importantes, bien délimitées, hétérogènes au maximum postérieur de part et d'autre des deux fémurs. Il existe une visualisation inflammatoire également des deux genoux, un peu plus nette que sur le reste des articulations, de même que sur les deux chevilles. Sur les mollets, il existe une discrète hétérogénéité postéro- interne.

CONCLUSION:
Fixation musculaire hétérogène et bien caractéristique,indéniable avec entésopathies associées sur les chevilles, les genoux, les hanches et les épaules.
Il est bien difficile de faire la part entre la responsabilité de la maladie musculaire et la responsabilité des entésopathies qui sont constantes chez ces patients.
Merci de me tenir au courant, cordialement.

Docteur N. CAILLAT-VIGNERON

Patient diabétique ?

Fumeur ?

Est capable de prendre ses propres décisions ?

Implique une décision/prédiction au niveau du patient, c'est-à-dire de plusieurs documents.

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft*

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Différentes tâches

Phénotypage simple: comment ?

1. Règles
   présence de mots-clés ou de motifs dans certaines parties de documents

1. Apprentissage supervisé
   classification de texte multi-classe / multi-label

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft*

scai
SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Différentes tâches

Phénotypage complexe

MAMMOGRAPHIE :

On décèle la présence de foyers de micro-calcifications (LÉSION)
dans le sein droit (LOCALISATION) dans le rayon de 8h (LOCALISATION), et dans
le sein gauche (LOCALISATION) dans le rayon de 7h (LOCALISATION).

CONCLUSION :

L'examen est donc reclassé ACR 4 (SCORE) de chaque côté (LOCALISATION).

SCORE ACR	LATÉRALITÉ
4	DROIT
4	GAUCHE

CIBLE	LATÉRALITÉ	QUADRANT	TAILLE
MICRO-CALCIF	DROIT	8h	?
MICRO-CALCIF	GAUCHE	7h	?

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft*

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Différentes tâches

Phénotypage complexe: comment ?

Ad-hoc selon les projets. Pas d’approche standard existante

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

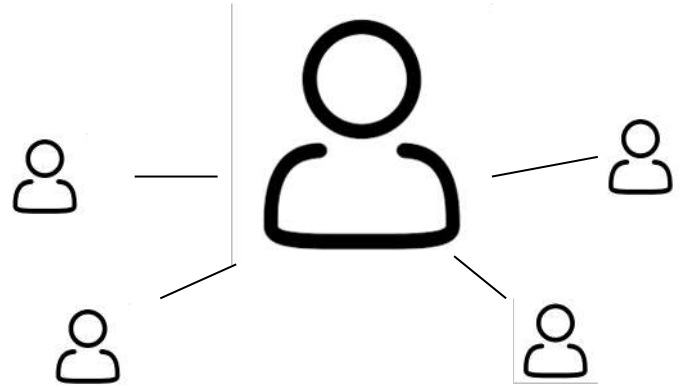
datacraft*

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Pourquoi ?

Soin

- Accès plus rapide à l'information
- Visualisation des parcours de soin
- Aide au diagnostic
- Recherche de cas similaire
- Comparaison de parcours de soin avec des protocoles standard



ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

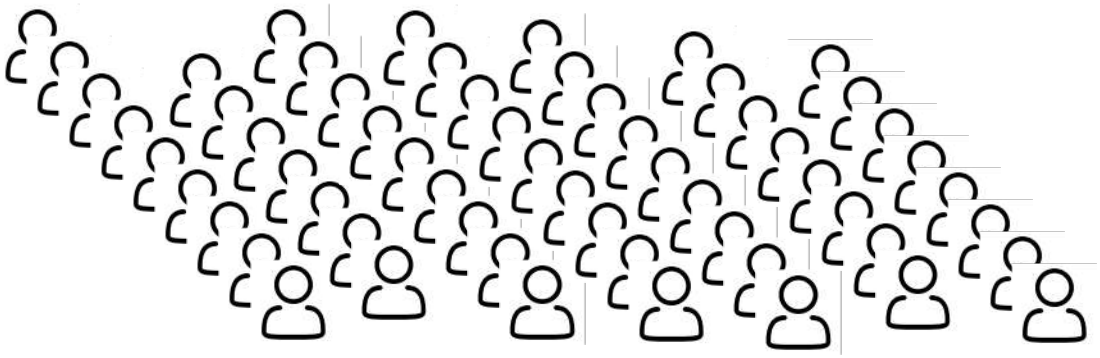
datacraft*

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Pourquoi ?

Recherche médicale, santé publique

- Désidentification des dossiers
- Sélection de cohortes pour des essais cliniques
- Études statistiques, fouille de données, …
- e.g. détection d'effets indésirabes de médicaments et de leurs causes

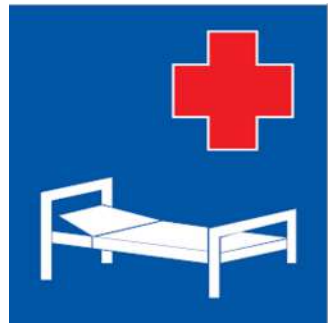


ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft*

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Pourquoi ?



Pilotage

- Codage medico-économique
- Organisation de l’hôpital

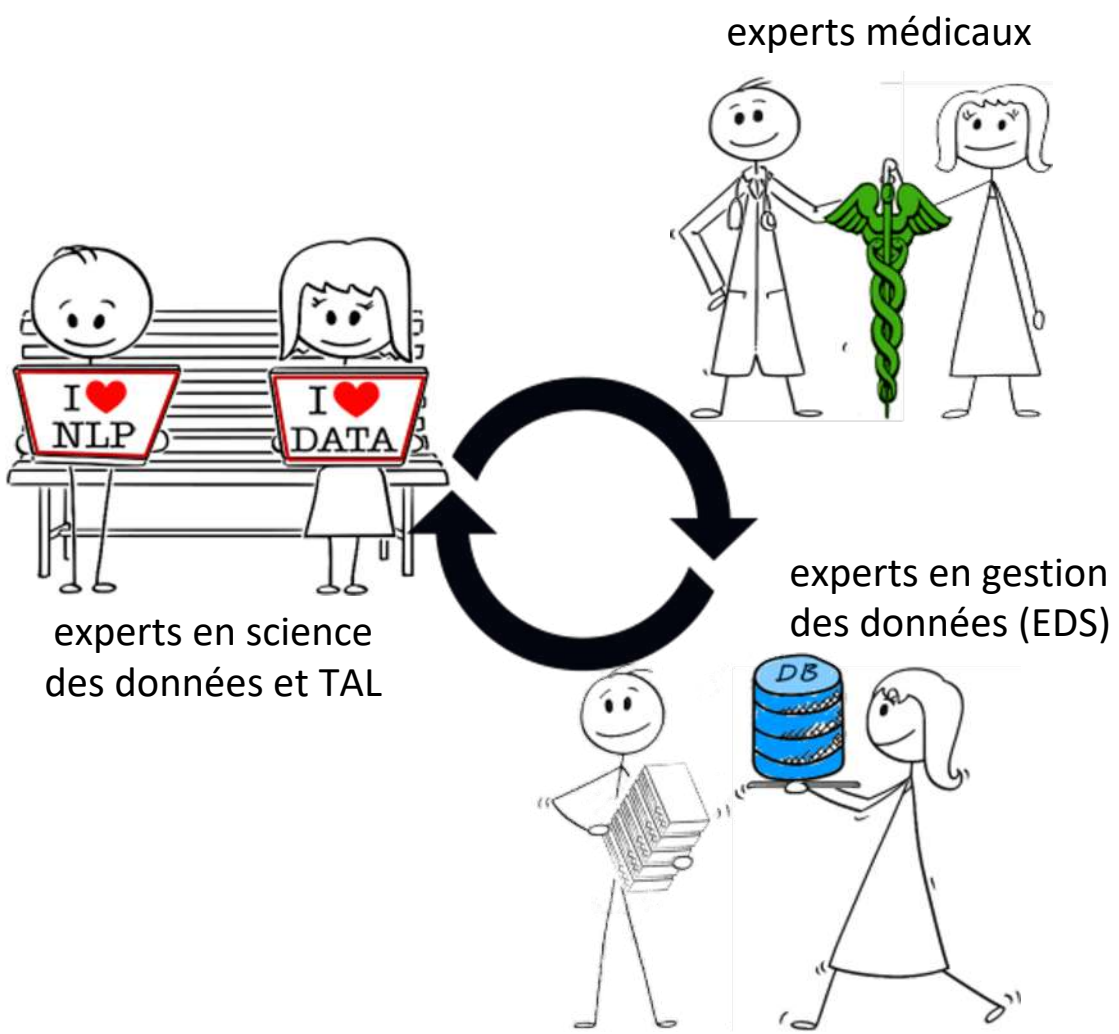
ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft*

scai
SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Protocole standard

Élaboration d'un protocole de discussion et de travail pour les problèmes qui relèvent d'approches TAL « classiques » et bien délimités



ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

2. Focus sur la normalisation

Perceval Wajsbürt

Objectif : apparier les entités nommées avec une base terminologique de concepts médicaux

Enjeux :

- **beaucoup** de concepts possibles (plusieurs centaines de milliers/millions)
- **peu** d'exemples par concepts (en moyenne **2 ou 3**)
- seulement **3.5%** des concepts ont un synonyme français, la plupart sont en anglais
- le modèle doit être rapide

[C0018674] (Craniocerebral Trauma)

Patient sans signe clinique évident de traumatisme crânien

Patiente avec antécédent de chirurchie bariatrique [C1456587] (Bariatric Surgery)

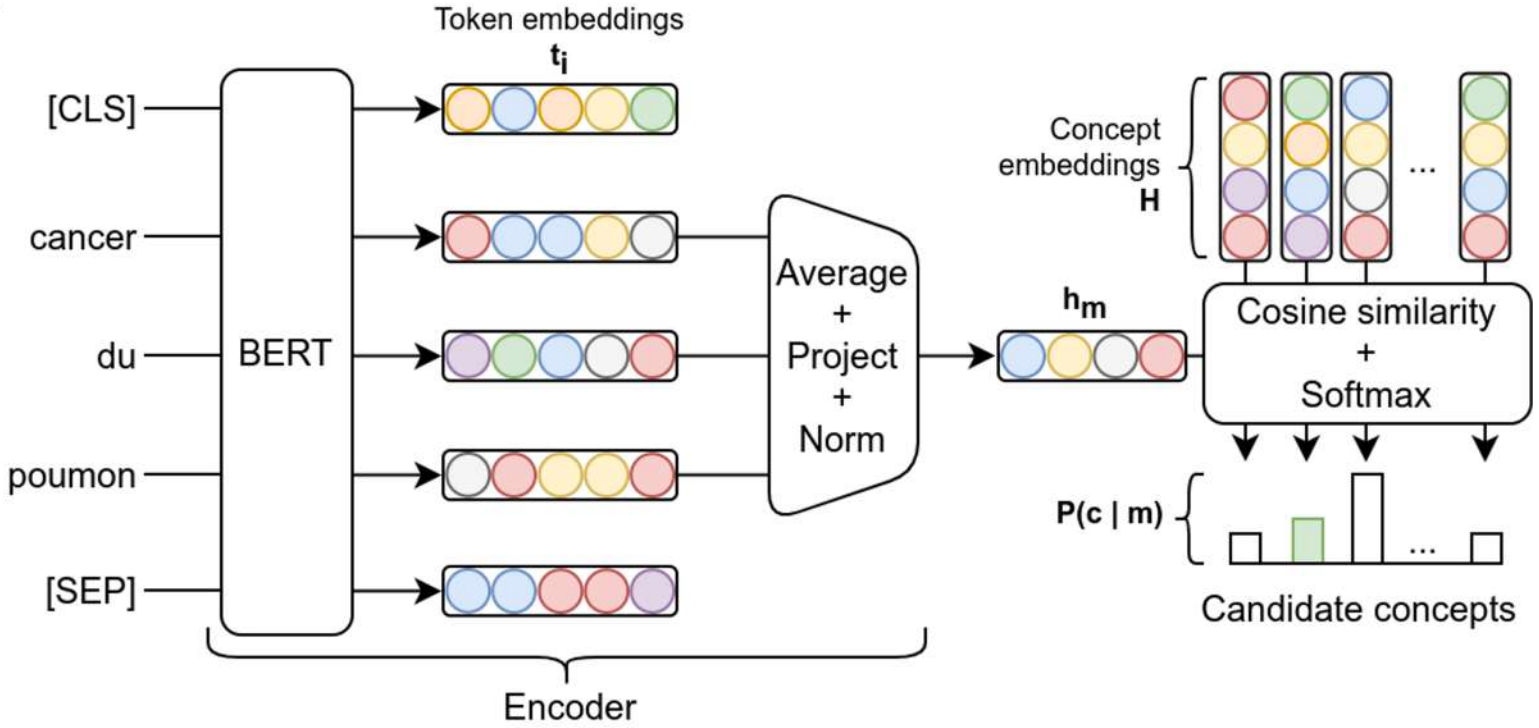
ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft*

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Focus sur la normalisation

Approche : classification à partir d'un modèle (BERT) pré-entrainé sur des données multilingues



ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft*

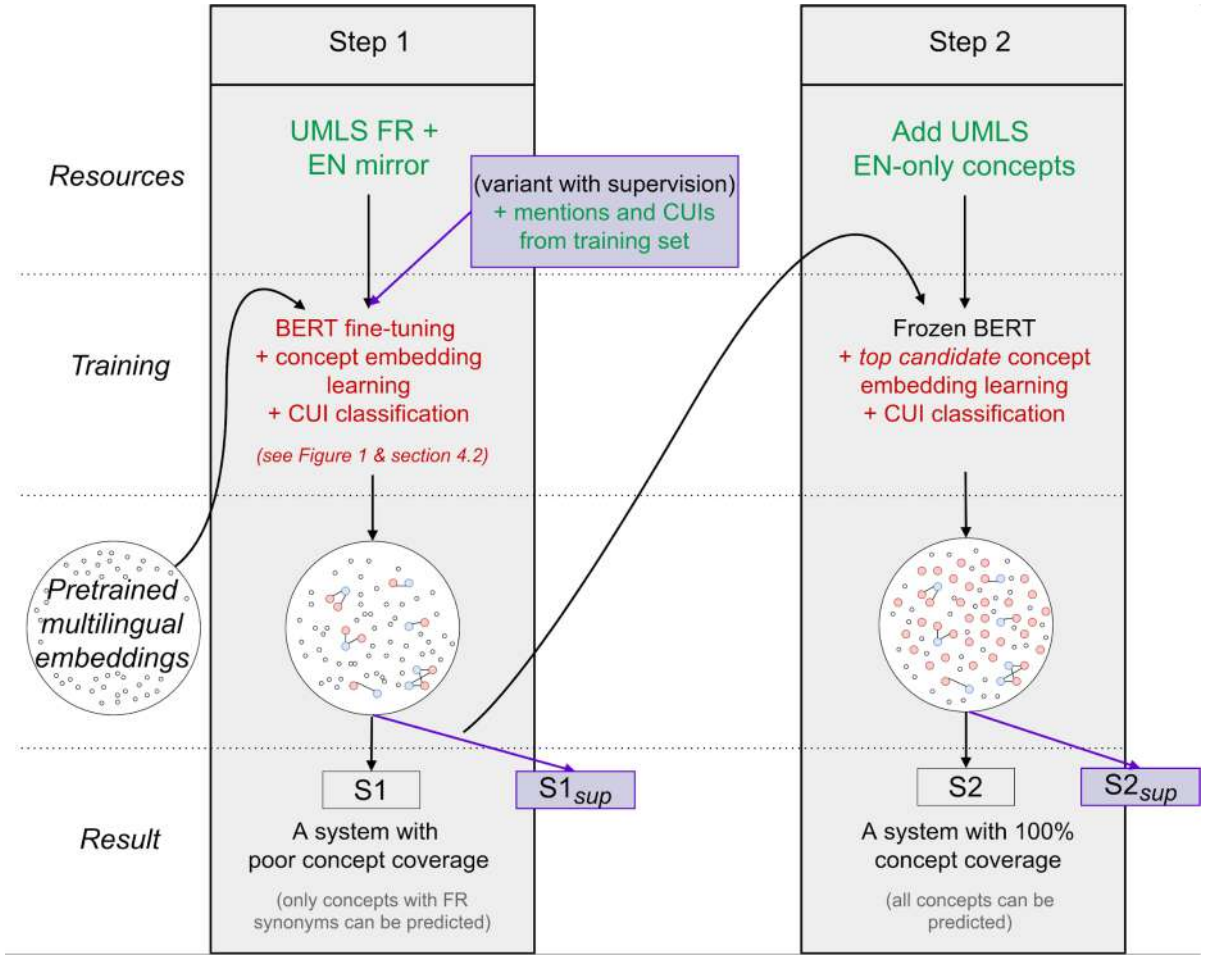
scai
SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Focus sur la normalisation

Entrainement:

- entraînement de tout le modèle (*encodeur* + *classifieur*) sur les concepts **multilingues**

- puis fine-tuning du *classifieur* seulement sur tous les concepts



ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft*

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Focus sur la normalisation

Résultats:

- Résultats très bons, même sans corpus annoté

	MEDLINE 2015	EMEA 2015	MEDLINE 2016	EMEA 2016
	F1	F1	F1	F1
Meilleur système CLEF 2015 (avec corpus annoté)	0.671	**0.872**	0.552	0.524
SOTA: traduction + recherche de terme (avec corpus annoté)	0.736	0.835	0.713	0.734
Supervision distante (sans corpus annoté)	0.737	0.765	0.754	0.727
Supervision (avec corpus annoté)	**0.790**	0.851	**0.795**	**0.743**

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

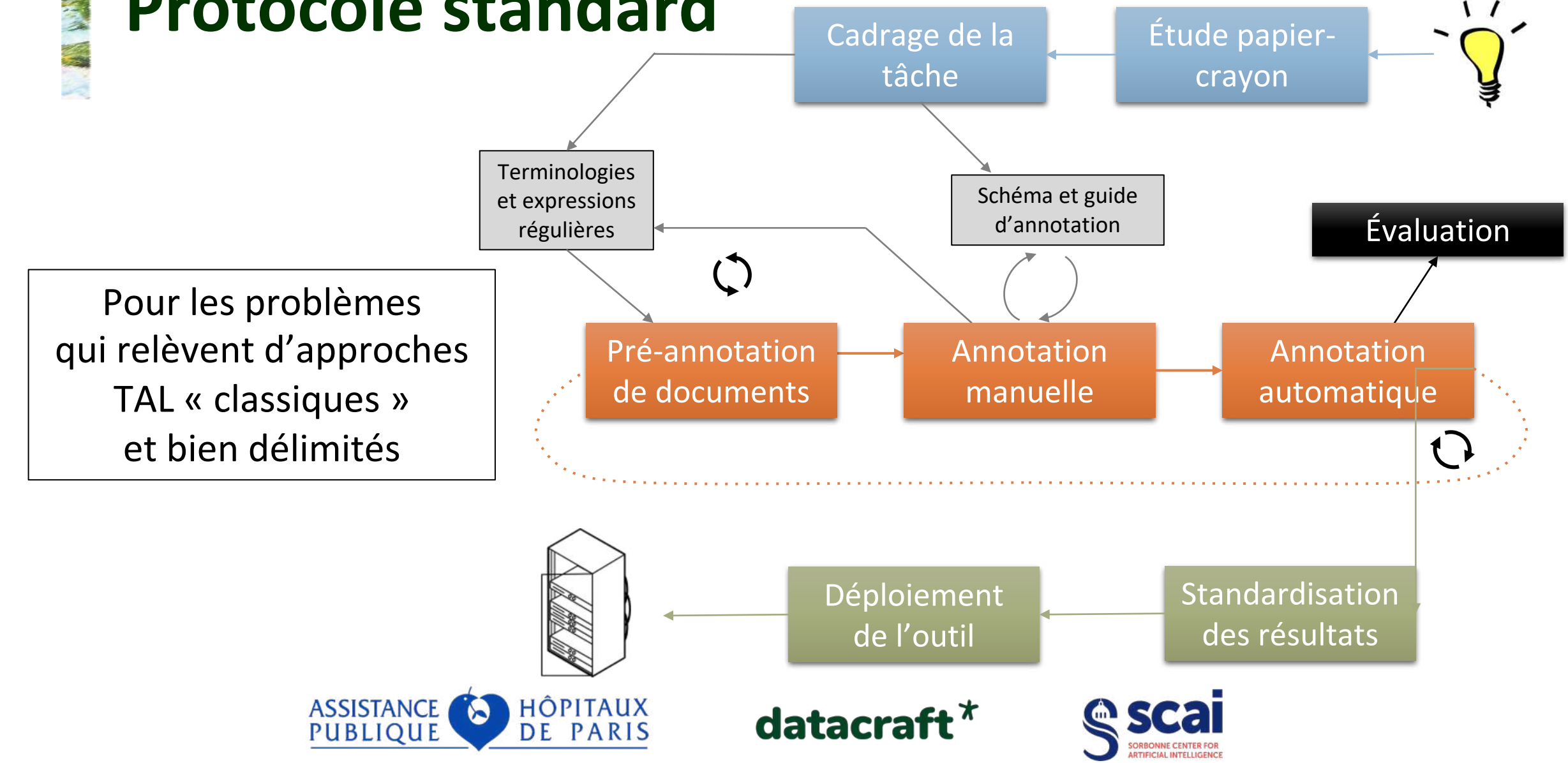
datacraft*

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Merci

Protocole standard

Pour les problèmes
qui relèvent d'approches
TAL « classiques »
et bien délimités



ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft*

scai
SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Focus sur la normalisation

Résultats:

- Résultats très bons, même sans corpus annoté

	MEDLINE 2015			EMEA 2015			MEDLINE 2016			EMEA 2016		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Meilleur système CLEF 2015 (avec corpus annoté)	0.805	0.575	0.671	**1.000**	0.774	**0.872**	0.594	0.515	0.552	0.604	0.463	0.524
SOTA: traduction + recherche de terme (avec corpus annoté)	**0.831**	0.661	0.736	0.909	0.772	0.835	0.771	0.663	0.713	0.781	0.692	0.734
Supervision distante (sans corpus annoté)	0.756	0.719	0.737	0.797	0.736	0.765	0.775	0.734	0.754	0.746	**0.709**	0.727
Supervision (avec corpus annoté)	0.806	**0.775**	**0.790**	0.875	**0.827**	0.851	**0.860**	**0.740**	**0.795**	**0.832**	0.670	**0.743**

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

datacraft

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE