Défi Fouille de Textes 2020

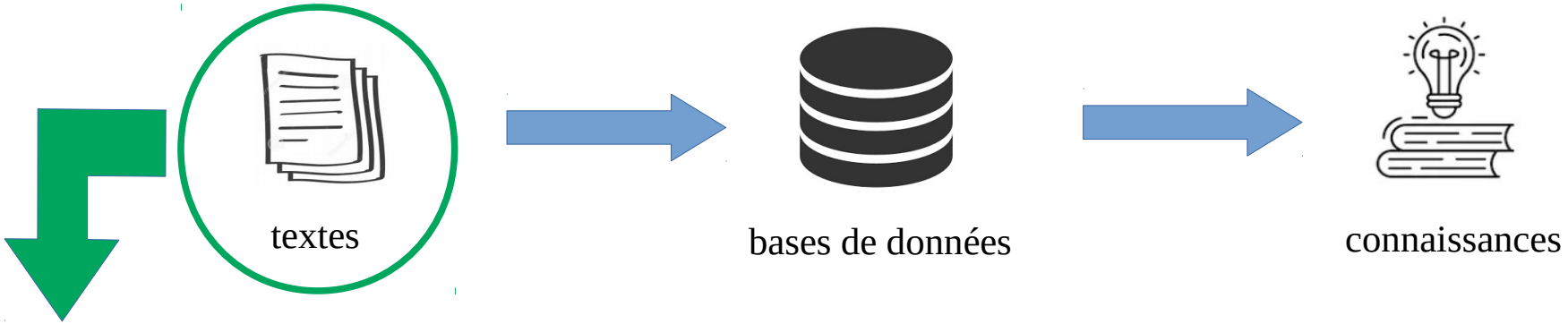
Cascade de CRF pour l'annotation d'entités cliniques imbriquées

Anne-Lyse Minard(2) **Andréane Roques**(1) **Nicolas Hiot**(1)
Mirian Halfeld Ferrari Alves(1) **Agata Savary**(3)

(1) Université d'Orléans, LIFO, Orléans, France
(2) Université d'Orléans, LLL-CNRS, Orléans, France
(3) Université François Rabelais Tours, LIFAT, Tours, France

Contexte : projet DOING

- **Activités** de **DOING** :



- **DEFT 2020 :**

  ⇒ **domaine médical** = première **cible d'application** des activités de DOING

  ⇒ **extraction d'information = aspect clé** du travail du groupe DOING

  ⇒ opportunité de **concrétiser une collaboration** dont l'ambition repose entre autres sur le **peuplement d'une base de données à partir des données textuelles**

Sommaire

1. Campagne d’évaluation DEFT 2020
   1.1 Présentation de la tâche 3 de DEFT
   1.2 Résumé de notre participation

2. Méthode

3. Système
   3.1 Pré-traitement
   3.2 Cascade de CRF
   3.3 Traits

4. Résultats

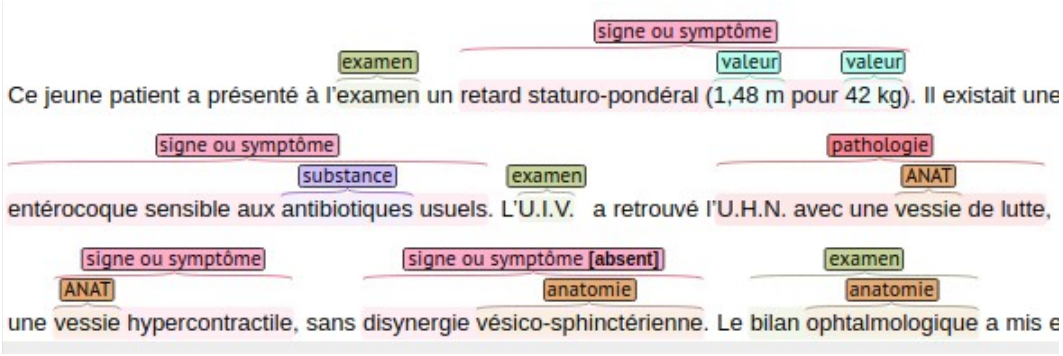
Conclusion

1.1 Présentation de la tâche 3 de DEFT (1)

- **Corpus :**
  - **cas cliniques** rédigés en **français**
  - **différentes spécialités médicales**
  - documents rédigés dans **plusieurs pays francophones**
  - **sources :**

    ⇒ CAS: French Corpus with Clinical Cases

    ⇒ Simple Corpus for Medical French (projet CLEAR)
  - **100 documents** pour le corpus d'apprentissage
  - **12 150 annotations** au total



- **Extraction d'informations fines autour de plusieurs catégories**

1.1 Présentation de la tâche 3 de DEFT (2)

- **Objectif de la tâche 3 =** dans des cas cliniques, extraire des informations fines autour de plusieurs catégories :
  - autour des patients : **anatomie**
  - autour de la pratique clinique : **examen, pathologie, signe ou symptôme**
  - autour des traitements médicamenteux et chirurgicaux : **substance, dose, durée, fréquence, mode d'administration, traitement, valeur**
  - autour du temps : **date, moment**

⇒ développer un système permettant d'**annoter automatiquement les entités** (anatomie, examen, pathologie, etc.) **et leurs informations** (valeur, mode, dose, etc.).

1.2 Résumé de notre participation

- Utilisation :
  - d'une **cascade de CRF** (champs aléatoires conditionnels) pour annoter les entités et leurs informations associées
  - d'une **ressource externe : MedDRA** (Dictionnaire Médical des Affaires Réglementaires)
  - d'un **pipeline (Ennov)**
- **3 runs soumis :**

	CRF	MedDRA	Pipeline Ennov	F-mesures (en moyenne, sur tous les types d'entités)
RUN 1	✓	✗	✗	**0,64**
RUN 2	✓	✓	✗	**0,65**
RUN 3	✓	✓	✓	**0,61**

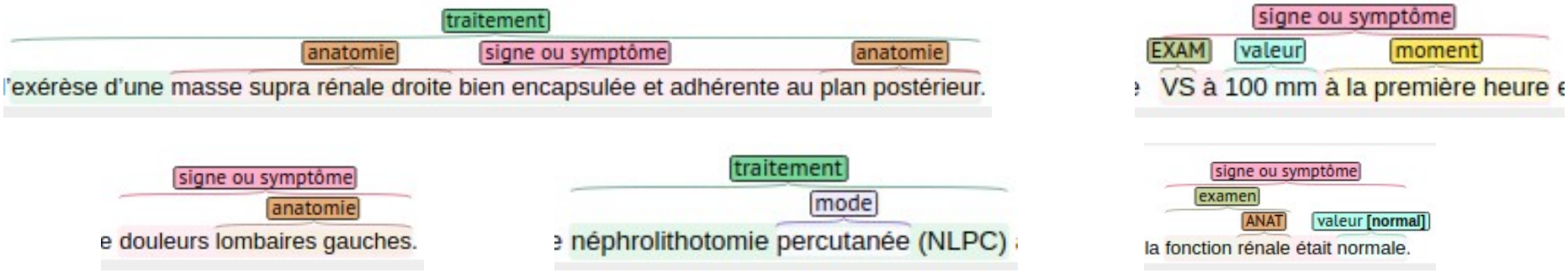
- **Problématiques et enjeux** principaux :

  ⇒ prise en compte de l'**imbrication des entités**

  ⇒ pertinence de l'**ordre d'enchaînement des entités** pour l'apprentissage en cascade

2. Méthode (1)

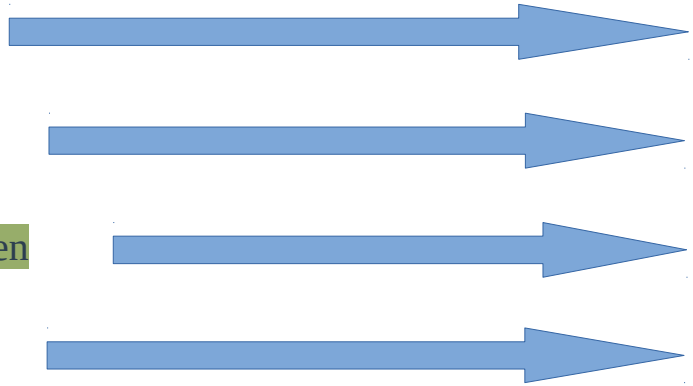
- **13 types d'entités** (dont 10 évalués) dans le **corpus d'entraînement** DEFT
- **2 groupes :**
  - **entités cliniques :** anatomie, examen, traitement, substance, sosy (signe ou symptôme), pathologie
  - **entités associées à ces entités cliniques :** valeur, dose, mode, moment (+ date, durée, fréquence)
- Choix de développer un **système d'apprentissage en cascade** étant donné les **nombreuses entités imbriquées** :



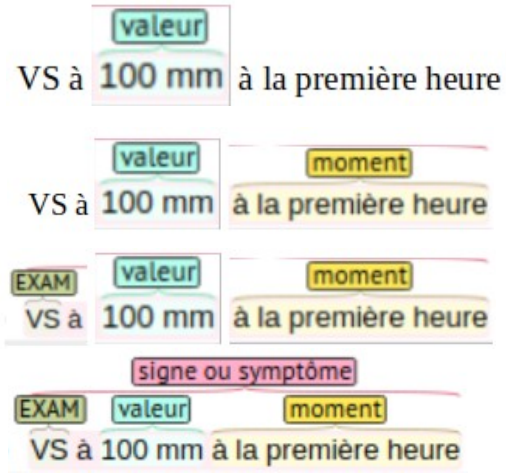
2. Méthode (2)

· **Apprentissage en cascade dans l'ordre suivant :**

1. dose et valeur
2. durée et fréquence
3. date et moment
4. anatomie et mode
5. traitement et examen
6. substance
7. sosy et pathologie



VS à 100 mm à la première heure



⇒ **regroupements d'entités proches sémantiquement, morphologiquement et qui ne sont pas imbriquées ou très peu**

Exemples de similarités morphologiques, sémantiques et/ou contextuelles :

- dose / valeur :

- moment / date :

⇒ **gains en rappel et/ou en précision** grâce à ces regroupements

⇒ permet de **prendre en compte** les **annotations des niveaux précédents** produites par le système

3. Système

- **Plusieurs modules** :

  1) module de **pré-traitement :**

  - **extraction des traits** (caractéristiques d'un token, d'un segment, etc. utiles pour apprendre à reconnaître les entités) utilisés ensuite pour l'apprentissage
  - transformation des **annotations** au **format BIO** (standard pour les CRF)

  2) **modèles appris pour chaque niveau d'annotation** : à partir des **fichiers templates** contenant des **traits** définis sous la forme de patrons, selon une syntaxe particulière

3.1 Pré-traitement

- **Outils et ressources** utilisés pour le pré-traitement des fichiers :
  - **sentence-splitter** pour le **découpage en phrases**
  - modèle français de **spaCy** (tokenisation, étiquetage morpho-syntaxique, etc.)
  - liste de **préfixes** et **suffixes** du français extraits du **TLFi**

- Extraits des **informations obtenues pour chaque token** :

TOKEN	LEMME	PoS	GENRE	NOMBRE	TYPE DE NUMÉRAL	FORME DU TOKEN	SUFFIXE	BIO-examen	BIO-sosy	BIO-valeur	BIO-moment
une	un	DET	Fem	Sing	-	xxx	False	O	O	O	O
VS	VS	PROPN	-	-	-	XX	False	B-examen	B-sosy	O	O
à	à	ADP	-	-	-	x	False	O	I-sosy	O	O
100	100	NUM	-	-	Card	ddd	False	O	I-sosy	B-valeur	O
mm	millimètre	NOUN	Masc	-	-	xx	False	O	I-sosy	I-valeur	O
à	à	ADP	-	-	-	x	False	O	I-sosy	O	B-moment
la	le	DET	Fem	Sing	-	xx	False	O	I-sosy	O	I-moment
première	premier	ADJ	Fem	Sing	Ord	xxxx	True - ière	O	I-sosy	O	I-moment
heure	heure	NOUN	Fem	Sing	-	xxxx	True - ure	O	I-sosy	O	I-moment

préfixes / suffixes = conservation du **plus long** :

« psych- » / « psycho- »

psychostimulant	True - psycho

« héma- » / « hémato- »

hématopoétiques	True - hémato

« lymph- » / « lympho- »

lymphonoeud	True - lympho

3.2 Cascade de CRF

- Outil **Wapiti** (Lavergne *et al.*, 2010)
- **Algorithme RPROP**
- **Paramètres par défaut** conservés

- **Prédictions faites pour chaque niveau conservées** dans le fichier de sortie du système
- **Prédictions sur les données non annotées** partiellement basées sur les **prédictions du/des niveau(x) précédent(s)**

3.3 Traits

- Définition des combinaisons de traits et des fenêtres suite à des **expérimentations en validation croisée à 10 plis :**

  - **Descripteurs sémantiques, morphologiques, morpho-syntaxiques** et **de surface**
  - **Bigrammes** : prise en compte de l'enchaînement des étiquettes choisies par le système
  - **Traits** issus de la **cascade CRF** : prise en compte des **annotations** des **niveaux précédents**

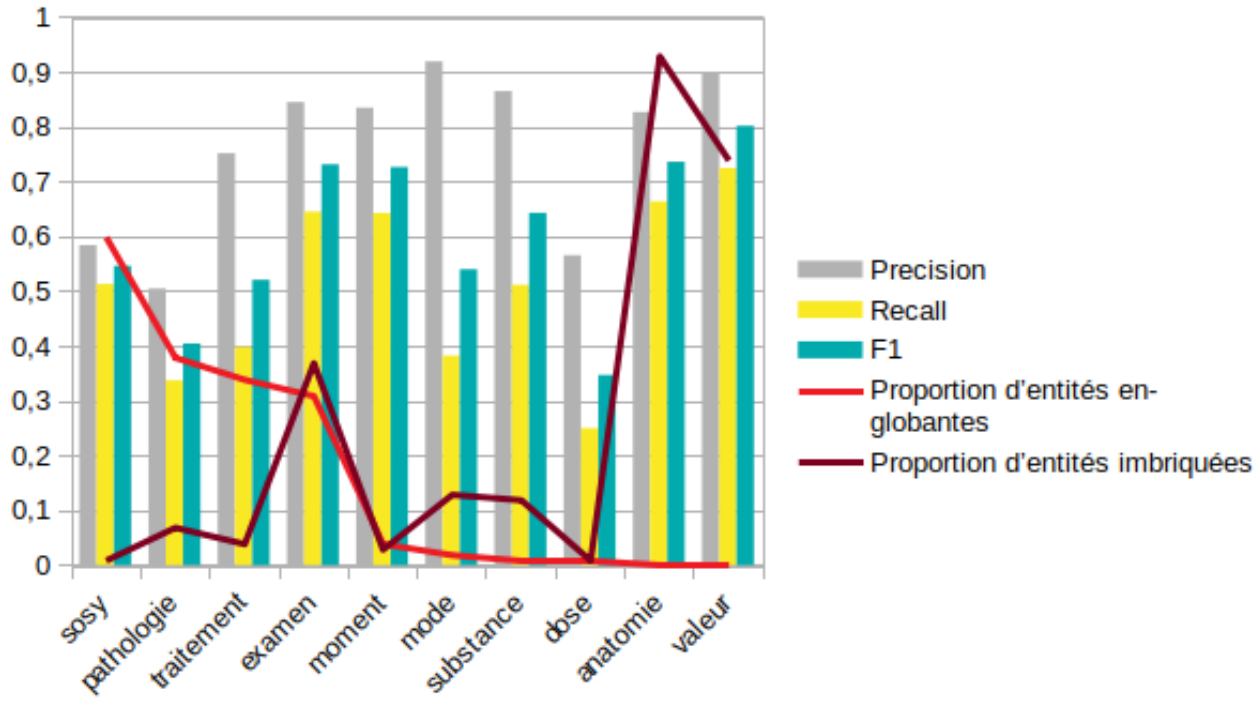
  - **Modèles** pour les **entités cliniques : traits** pour indiquer si un token fait partie d'une entité **MedDRA** (et le cas échéant, la classe associée) ou non
  - **Pipeline Ennov** : **traits** relatifs à la **position BIO du token dans les entités** reconnues par le pipeline (ce pipeline utilise également des CRF, mais ne prend pas en compte l'imbrication des entités)

4. Résultats (1)

	CRF	MedDRA	Pipeline Ennov	F-mesures (en moyenne, sur tous les types d'entités)
RUN 1	✓	✗	✗	**0,64**
RUN 2	✓	✓	✗	**0,65**
RUN 3	✓	✓	✓	**0,61**

⇒ meilleure F-mesure pour le run 2 = apports de MedDRA

4. Résultats (2)



(résultats du run 2)

- Entités **imbriquées** = plutôt **bien reconnues**
- Entités **longues et/ou englobantes** = **moins bien reconnues**

Conclusion

- Participation à la tâche d'**extraction d'informations fines** dans le **domaine médical**
- **Cascade** de **CRF** et gestion de l'**imbrication** des **entités**

- **Résultats supérieurs** à la **moyenne** et à la **médiane** de DEFT
- **Résultats de précision > résultats de rappel**

- **Pistes d'amélioration** envisagées :
  - **ajout d'informations lors du pré-traitement** (prise en compte du contexte, présence d'autres entités dans le contexte proche, etc.)
  - **généralisation** et **réduction** du nombre de **traits**

- Travail sur l'**extraction de relations entre des entités médicales**

Merci pour votre attention.

Des questions ?