Les données médicales dans les textes

Pierre Zweigenbaum[1]

LIMSI, CNRS, Université Paris-Saclay, Orsay, France

Congrès SIF, 7/2/2019

[1] Discussions avec Cyril Grouin, Aurélie Névéol, Thomas Lavergne, LIMSI

Plan

Plan

Plan

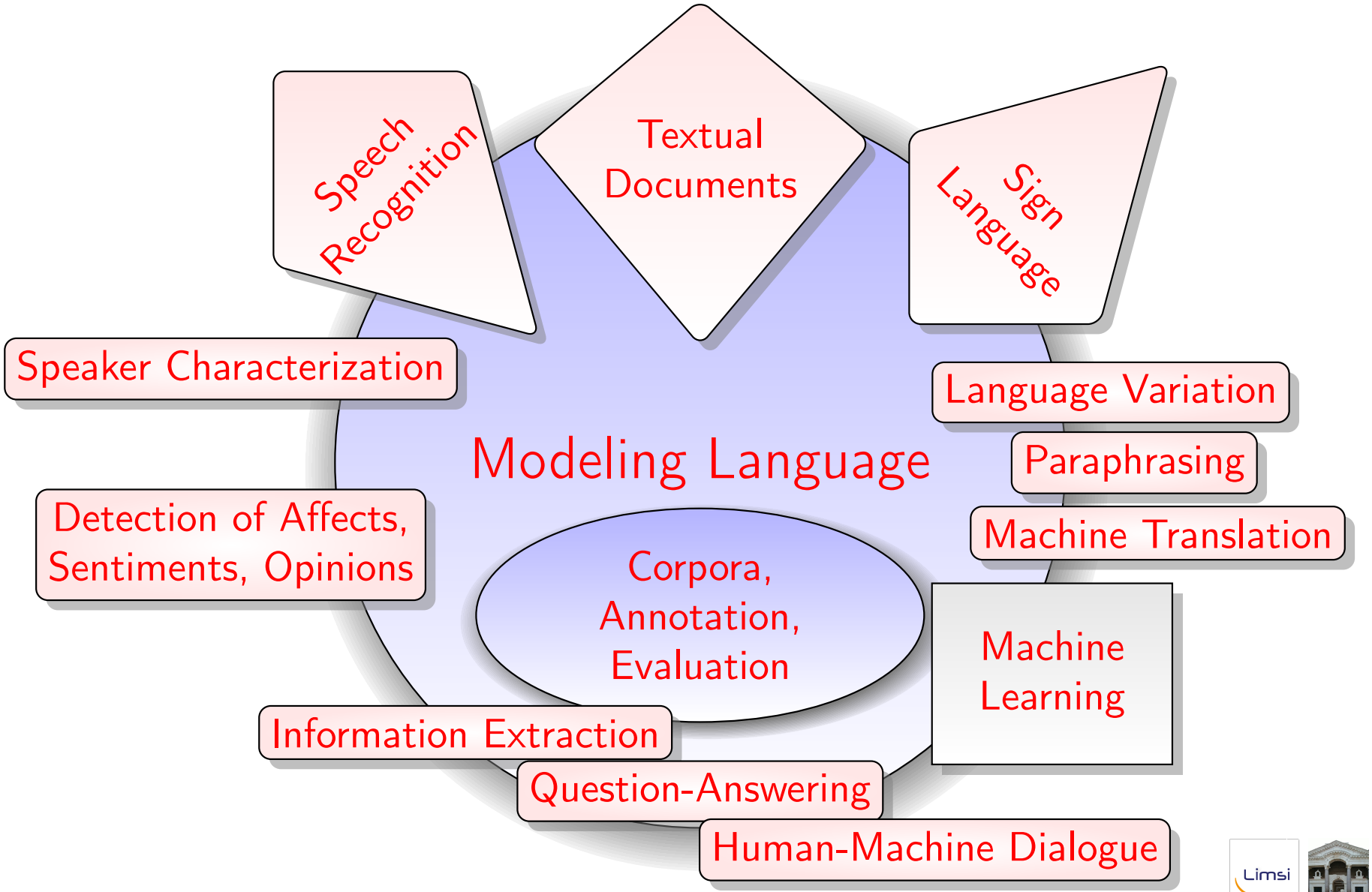
- Une partie importante des données et connaissances médicales est enregistrée et transmise sous forme de texte.
- L'obtention de ces données repose sur des méthodes de traitement automatique des langues.
- La mise au point de ces méthodes requiert des corpus de développement dont la constitution et le partage sont contraints par les impératifs de confidentialité des données médicales personnelles.
- Cela motive des travaux spécifiques sur la désidentification de ces textes et plus largement sur la conception d'autres méthodes de création de corpus de travail non confidentiels.
- Cela amène aussi à examiner la question du partage des modèles appris sur des textes à caractère confidentiel.

LIMSI: Un Laboratoire Interdisciplinaire pour la Mécanique et les Sciences de l'Information

Un laboratoire propre du CNRS sur le campus de l'Université Paris-Saclay

La nouvelle aile du LIMSI

Traitement automatique des langues au LIMSI



Plan

Objectifs : analyse de données textuelles

- Les textes médicaux portent et transmettent des informations et des connaissances
- L'analyse automatique de ces textes permet le traitement informatique de ces informations et connaissances

Bibliographie générale

- Névéol A, Zweigenbaum P. Expanding the Diversity of Texts and Applications: Findings from the Section on Clinical Natural Language Processing of the International Medical Informatics Association Yearbook. *Yearb Med Inform*, 27(01):193–198, 2018.
- Zweigenbaum P. Le traitement des langues naturelles dans le contexte de la eSanté. In Degoulet P, Fieschi M and Ménard J, e-Santé en perspective. Informatique et santé, vol. 20. Paris : Lavoisier, 2017.
- Demner-Fushman D, Elhadad N. Aspiring to unintended consequences of natural language processing: A review of recent developments in clinical and consumer-generated text processing. *Yearb Med Inform*. 2016 Nov 10;(1):224-233.
- Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*. 2008:128-44.
- Huang CC, Lu Z. Community challenges in biomedical text mining over 10 years: success, failure and the future. *Brief Bioinform*. 2016 Jan;17(1):132-44.
- Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. Frontiers of biomedical text mining: current progress. *Brief Bioinform*. 2007 Sep;8(5):358-75. Epub 2007 Oct 30.

Dossiers de patients

D'un texte libre...

Mon Cher Confrère,

...

Actuellement, sous Flécaïne 1 cp matin et soir et Préviscan, le patient est totalement asymptomatique. D'autre part, l'hypertension artérielle semble bien équilibrée par l'Aprovel 300, 1 par jour.

...

Au total, comme Monsieur Jean Dupont est actuellement peu symptomatique, je continuerai le même traitement sous la forme de Flécaïne 1 cp matin et soir en plus de l'Aprovel 300, 1 par jour. Par contre, je diminuerai progressivement le Préviscan et je le remplacerai par Kardégic 160 mg/24 h chez ce patient présentant une insuffisance aortique très modérée et une minime insuffisance mitrale sur prolapsus de la grande valve.

Dossiers de patients

D'un texte libre... à des données structurées : Ciblage d'entités et relations spécifiques

Mon Cher Confrère,

. . .

Actuellement, sous Flécaïne 1 cp matin et soir et Préviscan, le patient est totalement asymptomatique. D'autre part, l'hypertension artérielle semble bien équilibrée par l'Aprovel 300, 1 par jour.

. . .

Au total, comme Monsieur Jean Dupont est actuellement peu symptomatique, je continuerai le même traitement sous la forme de Flécaïne 1 cp matin et soir en plus de l'Aprovel 300, 1 par jour. Par contre, je diminuerai progressivement le Préviscan et je le remplacerai par Kardégic 160 mg/24 h chez ce patient présentant une insuffisance aortique très modérée et une minime insuffisance mitrale sur prolapsus de la grande valve.

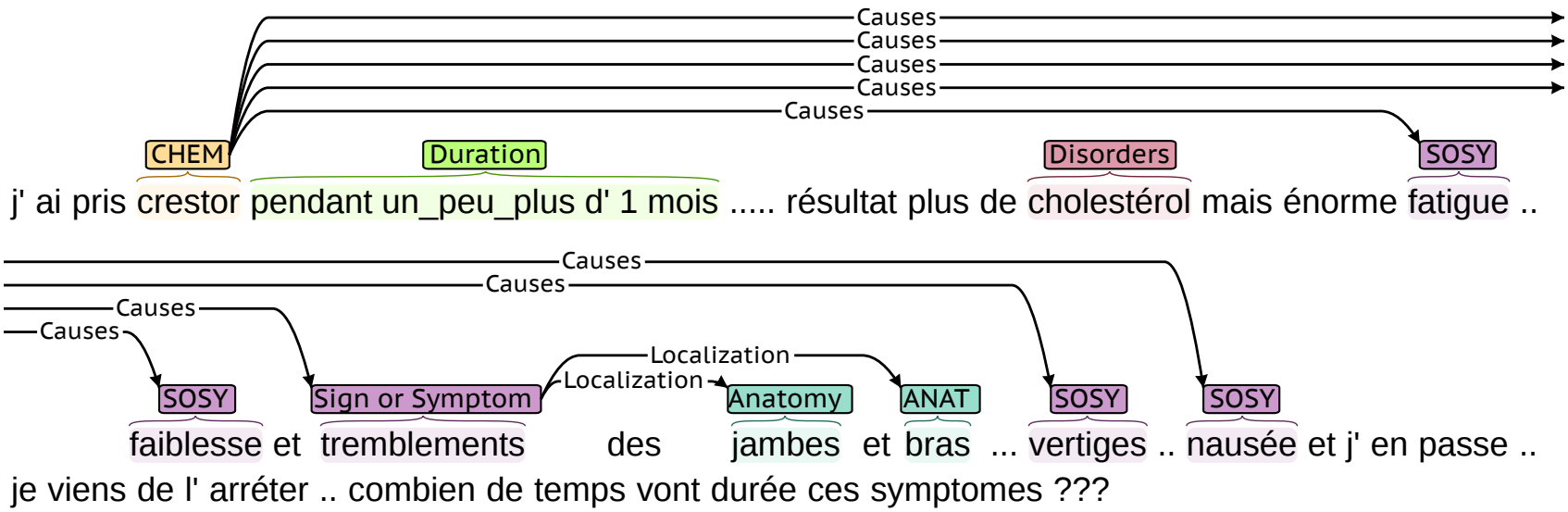
drug	dosage	frequency
flécaïne	1 cp	matin et soir
préviscan		
approvel 300	1	par jour
flécaïne	1 cp	matin et soir
approvel 300	1	par jour
préviscan		
kardégic	160 mg	/24 h

Étant donné un schéma prédéfini :

- entités
- relations

Forums de santé et pharmacovigilance

Trouver chaque mention d'un médicament, problème médical, etc., et leurs relations



Certificats de décès et statistiques

Repérer chaque cause de décès, trouver sa classe dans la Classification internationale des maladies

SEE INSTRUCTIONS ON OTHER SIDE

CAUSE OF DEATH

27. **PART I.** Enter the diseases, injuries, or complications that caused the death. Do not enter the mode of dying, such as cardiac or respiratory arrest, shock, or heart failure. List only one cause on each line.

Approximate Interval Between Onset and Death

IMMEDIATE CAUSE (Final disease or condition resulting in death) →

a. Rupture of myocardium — Mins.

DUE TO (OR AS A CONSEQUENCE OF):

Sequentially list conditions, if any, leading to immediate cause. Enter **UNDERLYING CAUSE** (Disease or injury that initiated events resulting in death) **LAST**

b. Acute myocardial infarction — 6 days

DUE TO (OR AS A CONSEQUENCE OF):

c. Chronic ischemic heart disease — 5 years

DUE TO (OR AS A CONSEQUENCE OF):

d.

PART II. Other significant conditions contributing to death but not resulting in the underlying cause given in Part I.

Diabetes, Chronic obstructive pulmonary disease, smoking

28a. WAS AN AUTOPSY PERFORMED? *(Yes or no)* — Yes

28b. WERE AUTOPSY FINDINGS AVAILABLE PRIOR TO COMPLETION OF CAUSE OF DEATH? *(Yes or no)* — Yes

29. MANNER OF DEATH

- [X] Natural
- [ ] Accident
- [ ] Suicide
- [ ] Homicide
- [ ] Pending Investigation
- [ ] Could not be Determined

30a. DATE OF INJURY *(Month, Day, Year)*

30b. TIME OF INJURY M

30c. INJURY AT WORK? *(Yes or no)*

30d. DESCRIBE HOW INJURY OCCURRED

30e. PLACE OF INJURY—At home, farm, street, factory, office building, etc. *(Specify)*

30f. LOCATION (Street and Number or Rural Route Number, City or Town, State)

Patient virtuel dialogant pour entraîner les étudiants

Discours, contexte

U— Avez-vous mal ?

S— Oui.

Patient virtuel dialogant pour entraîner les étudiants

Discours, contexte

U— Avez-vous mal ?

S— Oui.

U— À quel endroit ?

S— J'ai des douleurs de poitrine à droite.

U— Depuis quand ?

S— J'ai des douleurs depuis hier soir à 20 heures.

Le traitement automatique des langues

- Comment faire en sorte qu'un système informatique comprenne et produise du langage aussi bien que les humains

- *Natural Language Processing*

Intérêt du TAL dans le domaine médical

- Tâches de TAL
  - ∼ Convertit le texte en données
    - Extraction d'information
    - Indexation automatique, classification, codage
  - ∼ La langue comme moyen d'interaction
    - Système de dialogue personne-machine
  - ∼ Aide à l'usage humain de la langue
    - Dictée vocale
    - Correction automatique
    - Traduction automatique
- Usages
  - Statistiques, épidémiologie, santé publique
  - Accès aux connaissances, recherche d'information
  - Recherche médicale

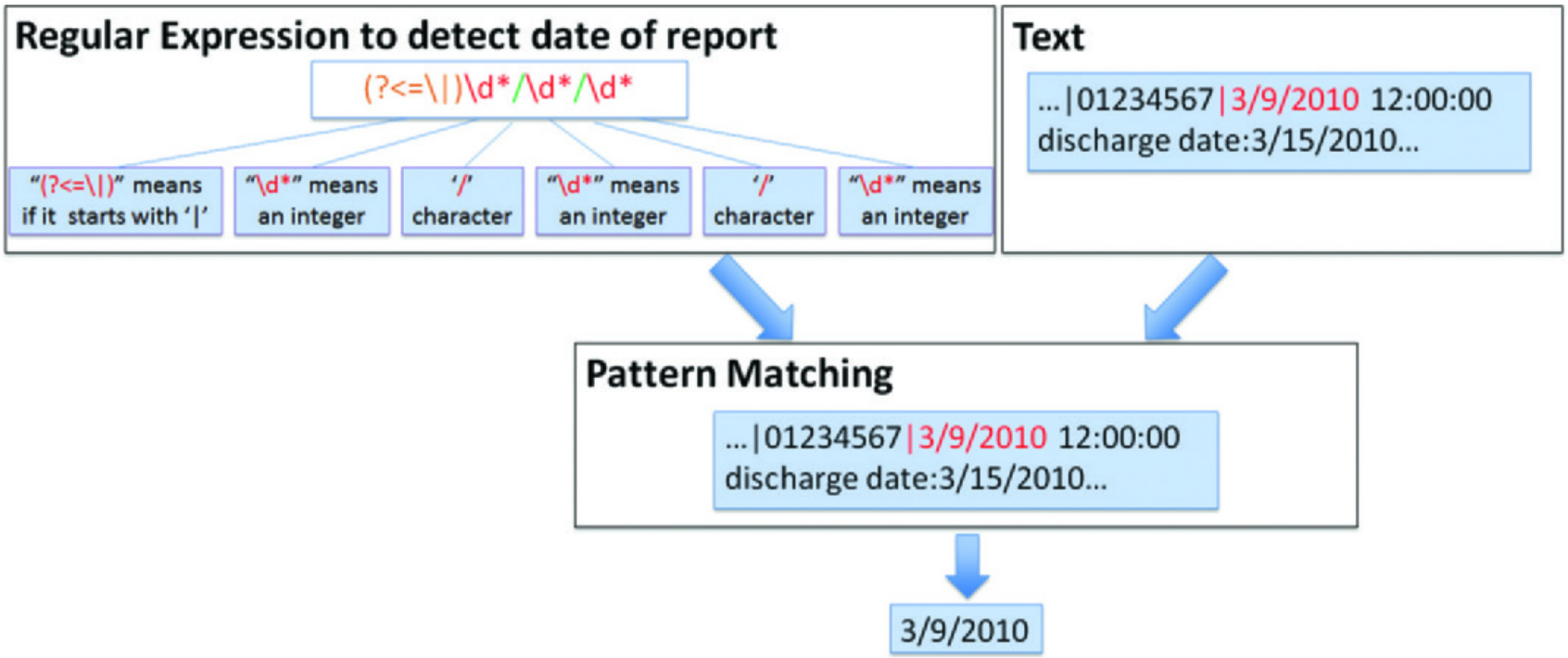
Plan

Problèmes

- Variation
  - infarctus du myocarde
  - infarctus myocardique
  - myocarde infarci
  - crise cardiaque
  - → On ne peut pas connaître tous les mots d'une langue
    - *cliniquement*, *cardiovasculaire*
    - noms de personnes, *Alzheimer*
    - *pt*, *SRAS*
    - *hémoragie*, *Alzeimer*
- Ambiguïté
  - iris
  - ventricule
  - genou
- Absence d'une spécification formelle complète
  - Langue naturelle $\neq$ Langage formel
- Besoin de connaissances
  - Fournies par des humains
  - Acquises par apprentissage automatique à partir de données

Méthodes à base de connaissances humaines

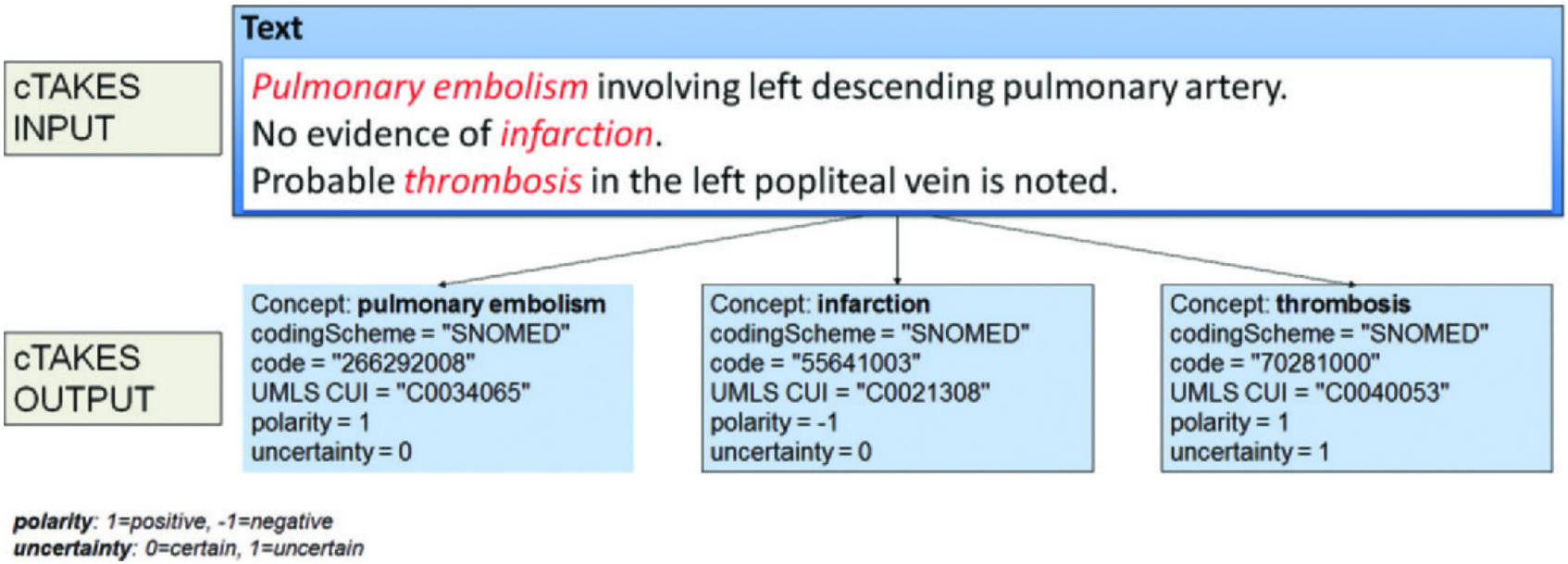
Exemple : expressions régulières



Cai T, Giannopoulos AA, Yu S, Kelil T, Ripley B, Kumamaru KK, et al. Natural Language Processing Technologies in Radiology Research and Clinical Applications. Radiographics 2016 Jan-Feb;36(1):176-91.

Méthodes à base de connaissances humaines

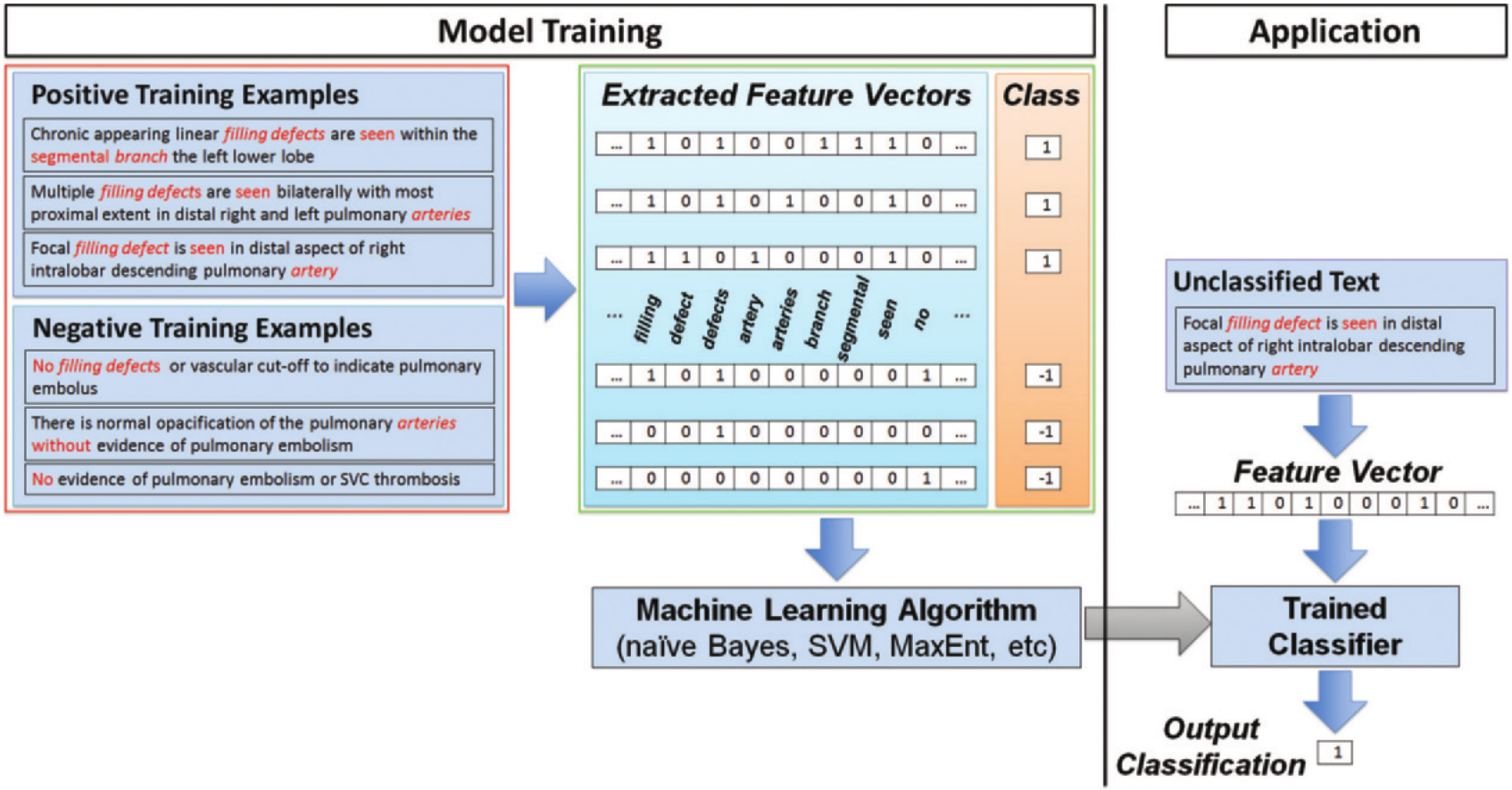
Exemple : usage d'un dictionnaire + détection de négation et modalité



Cai T, Giannopoulos AA, Yu S, Kelil T, Ripley B, Kumamaru KK, et al. Natural Language Processing Technologies in Radiology Research and Clinical Applications. Radiographics 2016 Jan-Feb;36(1):176-91.

Méthodes guidées par les données : apprentissage supervisé

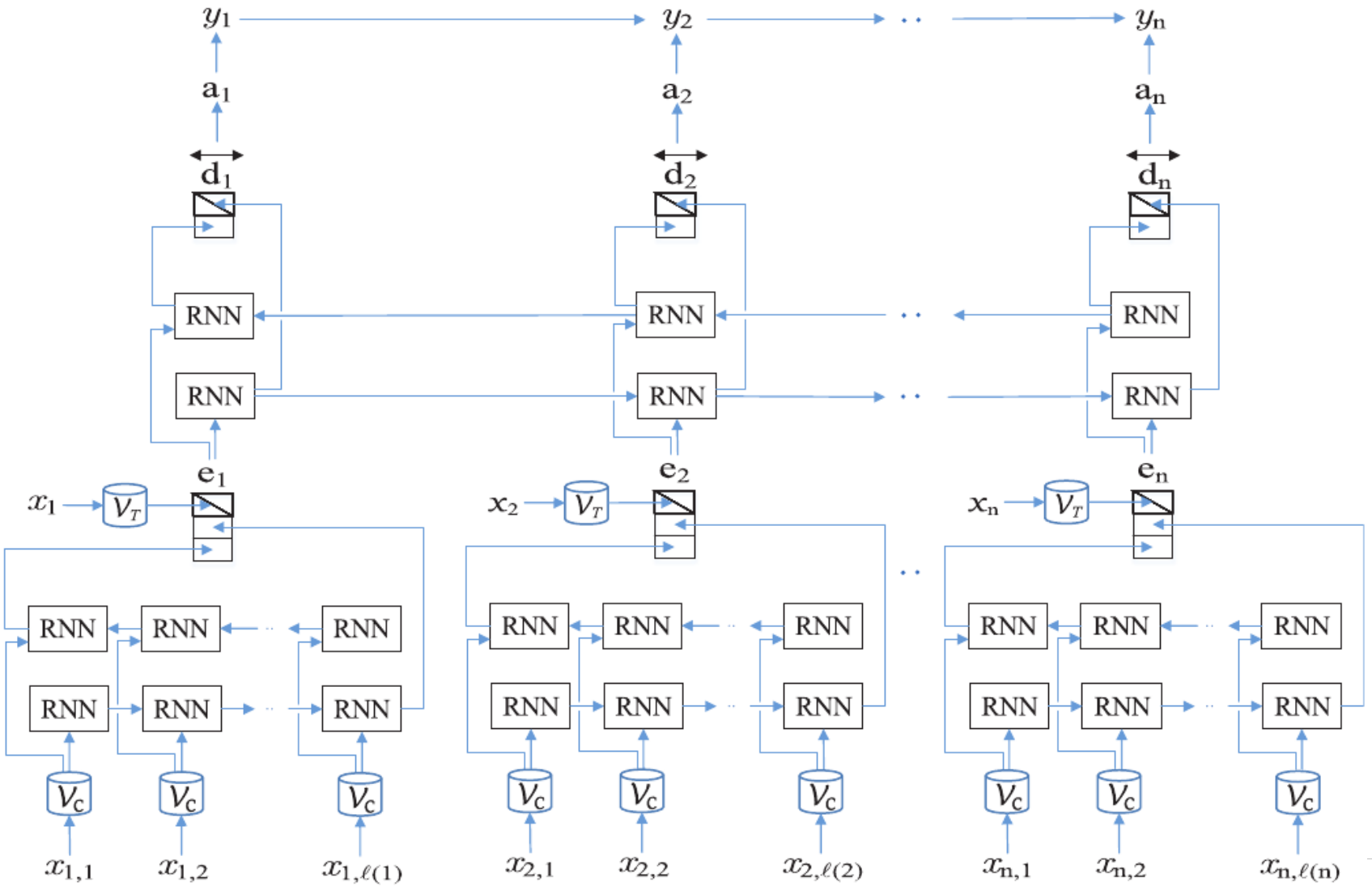
Classification de textes



Cai T, Giannopoulos AA, Yu S, Kelil T, Ripley B, Kumamaru KK, et al. Natural Language Processing Technologies in Radiology Research and Clinical Applications. Radiographics 2016 Jan-Feb;36(1):176-91.

Apprentissage supervisé : étiquetage de séquence

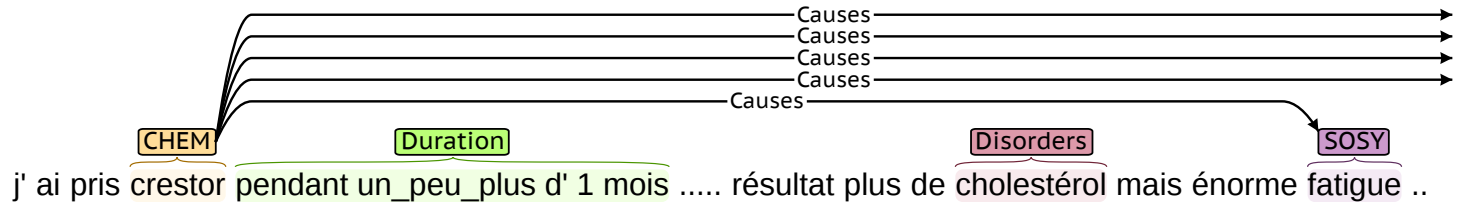
Réseaux de neurones récurrents sur mots et caractères, couche CRF

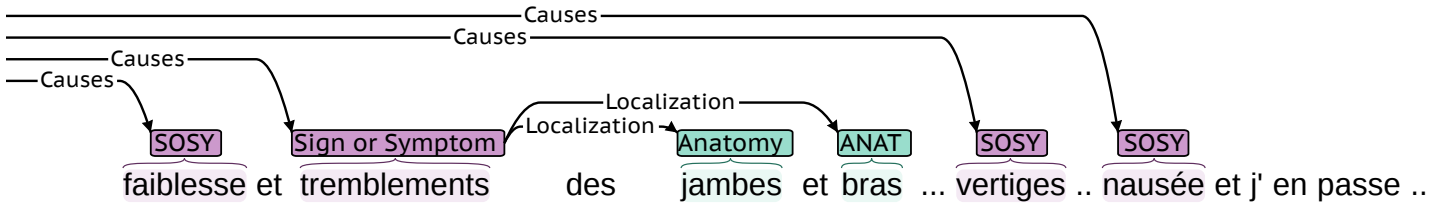


Lample et al. [2016], Dernoncourt et al. [2016]

Entraînement de méthodes par apprentissage

- Besoin de données annotées
  - Texte source
  - Vérité terrain ajoutée par un humain





  - je viens de l' arréter .. combien de temps vont durée ces symptomes ???
- Ces données peuvent être confidentielles
  - Dossier patient
- Restriction importante
  - pour la mise au point de méthodes et systèmes d'extraction d'information à partir de textes cliniques
  - nécessité d'anonymisation / désidentification

Plan

Besoin d'anonymisation d'un dossier de patient

Confidentialité

- Restriction aux professionnels de santé qui traitent le patient
- Besoin d'utilisation secondaire des données de santé

  "to realize the potentials for high quality healthcare, improved healthcare management, reduced healthcare costs, population health management, and effective clinical research"[2]
- Besoin plus large pour la recherche, notamment en TAL
  - Pour mettre au point diverses méthodes d'analyse
    - Pour aider la pratique médicale :
    - prise de décision
    - recherche de cas rares similaires
    - épidémiologie
    - pharmacovigilance
  - Notamment, des méthodes de désidentification !

[2] Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. Yearb Med Inform. 2017 Aug;26(1):38-52.

Un traitement en deux phases

- Détection d'entités
  - Détecte les frontières de chaque entité
  - Détermine le type de chaque entité
- Masquage des informations
  - Suppression des entités repérées
  - Remplacement par une étiquette générique : PATIENT, TÉLÉPHONE...
  - Remplacement par des substituts plausibles
    - Autres noms
    - Décalage des dates
  - Gestion de la cohérence des informations

Désidentification (MEDINA [Grouin and Zweigenbaum, 2013])

MEDINA – MEDical INformation Anonymization

Original text	Je revois ce 20 novembre 2012 Monsieur Jean Dumont (12.08.1924 ; 91 ans), qui a séjourné dans le service pour bilan du 10 au 12 février 2012. Ses principaux antécédents sont une BPCO, une gastrectomie pour ulcère de l'estomac il y a 30 ans (CHU Bordeaux) ; splénectomie en décembre 2008 ; une néphrectomie partielle gauche en janvier 2009 (Dr Rochelière) pour adénocarcinome d'évolution favorable ; une notion de canal lombaire étroit et une rupture de la coiffe des rotateurs de l'épaule droite (juillet 2007). M. Dumont reviendra le 3 mai 2012 pour contrôle. Ses enfants Jean-Pierre (67 ans) et Catherine (64 ans) s'interrogent sur la suite à donner. Dr. Gustave Le Dervant, 5 rue Jeanne Hachette, 75015 Paris - tél. 01 45 31 08 72 (sur rendez-vous)
Tagging	Je revois ce date 20 novembre 2012 Monsieur prenom Jean nom Dumont (date 12.08.1924 ; age 91 ans), qui a séjourné dans le service pour bilan du date 10 au 12 février 2012. Ses principaux antécédents sont une BPCO, une gastrectomie pour ulcère de l'estomac il y a 30 ans (hopital CHU Bordeaux) ; splénectomie en date décembre 2008 ; une néphrectomie partielle gauche en date janvier 2009 (Dr nom Rochelière) pour adénocarcinome d'évolution favorable ; une notion de canal lombaire étroit et une rupture de la coiffe des rotateurs de l'épaule droite (date juillet 2007). M. nom Dumont reviendra le date 3 mai 2012 pour contrôle. Ses enfants prenom Jean-Pierre (67 ans) et prenom Catherine (64 ans) s'interrogent sur la suite à donner. Dr. prenom Gustave nom Le Dervant, adresse 5 rue Jeanne Hachette, codepostal 75015 ville Paris - tél. telephone 01 45 31 08 72 (sur rendez-vous)
Date shifting	Je revois ce date 17 août 2011 Monsieur prenom Jean nom Dumont (date 09.05.1923 ; age 91 ans), qui a séjourné dans le service pour bilan du date 6 au 8 novembre 2010. Ses principaux antécédents sont une BPCO, une gastrectomie pour ulcère de l'estomac il y a 30 ans (hopital CHU Bordeaux) ; splénectomie en date septembre 2007 ; une néphrectomie partielle gauche en date octobre 2007 (Dr nom Rochelière) pour adénocarcinome d'évolution favorable ; une notion de canal lombaire étroit et une rupture de la coiffe des rotateurs de l'épaule droite (date avril 2006). M. nom Dumont reviendra le date 28 janvier 2011 pour contrôle. Ses enfants prenom Jean-Pierre (67 ans) et prenom Catherine (64 ans) s'interrogent sur la suite à donner. Dr. prenom Gustave nom Le Dervant, adresse 5 rue Jeanne Hachette, codepostal 75015 ville Paris - tél. telephone 01 45 31 08 72 (sur rendez-vous)
Pseudonymization	Je revois ce date 17 août 2011 Monsieur Charlie Martin (date 09.05.1923 ; age 91 ans), qui a séjourné dans le service pour bilan du date 6 au 8 novembre 2010. Ses principaux antécédents sont une BPCO, une gastrectomie pour ulcère de l'estomac il y a 30 ans (hopital CHU Bordeaux) ; splénectomie en date septembre 2007 ; une néphrectomie partielle gauche en date octobre 2007 (Dr Dubois) pour adénocarcinome d'évolution favorable ; une notion de canal lombaire étroit et une rupture de la coiffe des rotateurs de l'épaule droite (date avril 2006). M. Martin reviendra le date 28 janvier 2011 pour contrôle. Ses enfants Claude (67 ans) et Alex (64 ans) s'interrogent sur la suite à donner. Dr. Camille Bernard, adresse 5 rue Jeanne Hachette, codepostal 75015 ville Paris - tél. telephone 01 45 31 08 72 (sur rendez-vous)
	Je revois ce 17 août 2011 Monsieur Charlie Martin (09.05.1923 ; <age /> ans), qui a séjourné dans le service pour bilan du 6 au 8 novembre 2010.

Quelles informations supprimer

- Noms de personnes
  - nom de famille
  - prénom
  - nom de patient
  - nom de médecin ?
- Identifiants numériques
  - Téléphone
  - Courriel
  - Numéro de sécurité sociale
  - Numéro de dispositif médical
  - etc.
- Noms de lieux
  - Hôpital ?
  - Adresse
  - ...
- Autres informations
  - Métier, événements spécifiques

HIPAA Safe Harbor

18 types d'informations identifiantes

Table 1. PHI types as defined by HIPAA, i2b2, and MIMIC

PHI categories	PHI types	Descriptions	HIPAA	i2b2	MIMIC
AGE	AGE	Ages ≥90	x	x	x
		Ages <90		x	
CONTACT	PHONE	Telephone numbers	x	x	x
	FAX	Fax numbers	x	x	PHONE
	EMAIL	Electronic mail addresses	x	x	
	URL	Uniform resource locators	x	–	
	IP ADDRESS	Internet protocol addresses	x	–	
DATE	DATE	Dates (month and day parts)	x	x	x
		Year		x	x
		Holidays		x	x
		Days of the week		x	
ID	IDNUM	Social Security numbers	x	x	x
		Account numbers	x	x	x
		Certificate or license numbers	x	x	x
	MEDICAL RECORD	Medical record numbers	x	x	IDNUM
	DEVICE	Vehicle or device identifiers	x	x	IDNUM
	HEALTH PLAN	Health plan numbers	x	–	IDNUM
	BIOID	Biometric identifiers or full-face photographs	x	–	
LOCATION	STREET	Street address	x	x	x
	CITY	City	x	x	LOCATION-OTHER
	ZIP	Zip code	x	x	x
	STATE	State		x	x
	COUNTRY	Country		x	x
	LOCATION-OTHER	Other identifiable locations such as landmarks		x	x
	ORGANIZATION	Employers	x	x	
	HOSPITAL	Hospital name		x	x
		Ward name			x
NAME	PATIENT	Names of patients and family members	x	x	x
	DOCTOR	Provider name		x	x
	USERNAME	User IDs of providers		x	
PROFESSION	PROFESSION	Profession		x	

Classification of PHI into categories and types are as defined in the i2b2 dataset. During training, the PHI types are used as the labels to predict. The mark "–" denotes that two or fewer instances of the corresponding PHI types are present in the whole dataset, and no instance is present in the test set. In the MIMIC data-

Permormance de la désidentification automatique

Réseaux de neurones sur mots et caractères : résultats [Dernoncourt et al., 2016] sur les mêmes types de textes

	i2b2			MIMIC		
Model	Precision	Recall	F1	Precision	Recall	F1
Nottingham	**99.000**	96.400	97.680	–	–	–
MIST	91.445	92.745	92.090	95.867	98.346	97.091
CRF	98.560	96.528	97.533	99.060	98.987	99.023
ANN	98.320	97.380	97.848	**99.208**	99.251	**99.229**
CRF + ANN	97.920	**97.835**	**97.877**	98.820	**99.398**	99.108

Difficultés

- Forme variable, hétérogénéité
  - Noms
    - Avec ou sans prénom, initiales
    - Orthographe pas toujours respectée
  - Dates
    - Absolues ou relatives, précisant ou pas le jour, le mois, l'année
    - Format numérique ou en toutes lettres, abréviations
  - Téléphone
    - Sur 10 chiffres, extension seule
- Noms de personnes dans d'autres expressions
  - hôpital Robert Debré
  - salle Castaing
  - rue Ambroise Paré
  - talon d'Achille, maladie d'Alzheimer
- Différentes zones de texte
  - Parties structurées (en-tête de courrier ou de compte rendu)
  - Parties narratives (corps d'une lettre ou d'un rapport)
  - Colonnes dans un document « imprimé »

Difficultés par type [Dernoncourt et al., 2016]

Table 6. Examples of PHI instances undetected by CRF + ANN (i.e., undetected by both CRF and ANN) for the i2b2 dataset

PHI categories	PHI type	Examples	Reason	FN	Support
AGE	AGE	A **seventy-one**-year-old woman with multiple medical	S	19	790
		died of sudden death in their **82nd** year. Brother had SCD at **66**.	S		
		smoked from age 7 to 15, has not smoked since **15**.	S		
		d 80s?cause, MGF d**90** age, MGM d**73** CVAM d 73	S		
		stomach Ca, OA, obeseF d **84** multi-infarct dementiaS b66	S		
CONTACT	PHONE	Wheatland Manor: 154-734-1487, x**557** (4th floor)	S	1	410
	FAX	Phone: (091)920-5569 Fax: **(251)628-xxxx**	S	3	6
	EMAIL	E-Mail: **iparedes@oachosp.org**	S	3	3
DATE	DATE	PARONYCHIAL INFECTION: LEFT HAND **78**||Ectopic pregnancy: **74**	Am	60	12534
		alb 4.2|fe 50, tibc 204, ferritin 878 **8/27**|inr 1.1|pth 115 8/27	Am		
		Prior HDL 19.**8/67** TC 170, TG 162, H40, L98	Am		
		Referral submitted to GI**6/65**: saw GI - going for scope to eval pancreas	Am		
		DMSon b93D b94 GC due**22**D Fran b03 Abn	S		
		last seen in clinic in ||**11-70** after which time she left for	S		
ID	IDNUM	Influenza vaccine || Received 11/95 **MLL**||||	Am	9	382
		disp #100 order number **38/48**||ALLERGY||NKDA	S		
	MEDICALRECORD	Patient: Vincent Ware (71417347 **2Y**)	S	1	732
	DEVICE	Interrogation today of his Medtronic Kappa **QQ 626** pacemaker	S	4	12
LOCATION	STREET	–		0	416
	CITY	Oriented to "LCC" in "**Galena**," "March 2095." Speech fluent in Dutch.	S	8	344
	ZIP	–		0	144
	STATE	BP has been well-controlled in **VA**, usually in the 128 systolic range.	Ab/Am	9	205
	COUNTRY	is here with her husband who is translating from **columbian**.	S	13	130
	LOCATION-OTHER	travel hx to the Rockefeller Centre, more recent **global** travel	D	12	20
		and has infrequently visited **Storting** and **Acropolis**.	S		
	ORGANIZATION	diabetes diet - he enjoys a blueberry muffin from **RR Donnelley** daily.	S	42	147
		his level of fatigue. He continues to go to **the library** daily. He continues	D		
	HOSPITAL	were placed at Pomeroy Care Center (Big Rapids, **AC**) and also he	Ab/Am	44	1595
		Medication List for QUICK,ISABELLE Y 6557545 (**ATCH**) 52 F	Ab		
		2. DM, stable, Glyburide increased at **MS**. Dietary rec's||reviewed.	Ab/Am		
NAME	PATIENT	DMSon b93D b94 GC due22D **Fran** b03 Abn pap24 Nephropathy 3/25	Am	6	1450
		(HCP, daughter) 625-248-3647; **Flowers** (son) 705-690-8475	Am		
		Patient Name: JIMENEZ,YOUSSEF **I** [0554733(LCH)]	Ab/Am		
	DOCTOR	Insley/Endocrinology - End 6||**Lane**/Neurology - NEU 265	Am	35	3297
		Script: Amt: 30 Refill: 3 Date: 03/11/2074: **um**	Am		
		If the latter, will change it.||O||||Plasma Sodium 138	Ab/Am		

Un corpus désidentifié disponible

- MIMIC II, MIMIC III [Saeed et al., 2011]
  - Boston, Mass.
  - En anglais
  - Dossiers de patients en soins intensifs
- Mise à disposition pour la recherche
  - Contrat de mise à disposition contraignant

Plan

Risque de réidentification

- Déterminer de quel patient provient un compte rendu
  - Un médecin du service peut-il retrouver un patient[3]
- Déterminer si un patient donné est présent dans un corpus de comptes rendus
- Voir les études de Scaiano et al. (2016)[4]
  - Habituellement : micro-mesure sur les entités
  - Probabilité de fuite pour un document
    - Intervalle de confiance de 95 %
  - Identifiants directs
    - 0.0074 (0,002–0,016)
  - Quasi-identifiants
    - 0.0022 (0,000–0,013)

[3] Cyril Grouin, Nicolas Griffon, Aurélie Névéol. Is it possible to recover personal health information from an automatically de-identified corpus of French EHRs? Proc of LOUHI, 2015.

[4] Martin Scaiano, Grant Middleton, Luk Arbuckle, Varada Kolhatkar, Liam Peyton, Moira Dowling, Debbie S. Gipson, Khaled El Emam. A unified framework for evaluating the risk of re-identification of text de-identification tools. Journal of Biomedical Informatics 63 (2016) 174–183.

Viser un rappel parfait ?

Risque de suppression d'informations utiles, voire nécessaires

- Marqueurs biologiques
  - Nécessaires pour la recherche sur les maladies rares[5]
- Résultats de laboratoire, potentiellement identifiants en combinaison
  - Remplacement par des intervalles risque de fausser les interprétations futures
- Informations géographiques et temporelles
  - Utilies pour les recherches épidémiologiques[6]

---

[5] Hansson, Mats G et al. "The Risk of Re-Identification versus the Need to Identify Individuals in Rare Disease Research." European Journal of Human Genetics 24.11 (2016): 1553–1558.

[6] Mazumdar S, Konings P, Hewett M, Bagheri N, McRae I, Del Fante P. Protecting the privacy of individual general practice patient electronic records for geospatial epidemiology research. Aust N Z J Public Health. 2014 Dec;38(6):548-52.

Documents externes aux textes

- Ne pas retrouver le patient demande une désidentification multimodale
  - texte (comptes-rendus)
  - image (radiographies)
  - numérique (résultats de laboratoire)
- La réidentification suppose
  - L'accès aux bases de patients d'un hôpital
  - De savoir interroger l'outil[7]

[7] Cyril Grouin, Nicolas Griffon, Aurélie Névéol. Is it possible to recover personal health information from an automatically de-identified corpus of French EHRs? Proc of LOUHI, 2015.

Autres solutions

- Phrases récurrentes
  - Une phrase (une expression ?) répétée dans les dossiers de plusieurs patients n'est pas identifiante
- Traduction de textes désidentifiés
  - Base américaine MIMIC
  - Traduction automatique
  - Révision humaine
    - par traducteur : cohérence
    - par médecin : conformité aux pratiques locales
- Génération automatique de textes
  - Besoin d'entraînement sur une base de textes réels
  - Risque de transfert d'informations identifiantes
- Création de textes fictifs
  - Demander à des médecins de créer des comptes rendus
- Usage de cas publiés
  - Campagne d'évaluation DEFT 2019

« Modèles » appris sur des textes

- La phase d'entraînement d'un algorithme d'apprentissage apprend un « modèle »
- Un modèle enregistre des informations sur ses données d'entraînement
  - Caractéristiques (features)
- Exemple : « modèle de langue » n-gramme
  - *probabilités des n-grammes de mots*
  - probabilité de chaque mot
  - probabilité de chaque séquence de deux mots
  - ...
- L'information enregistrée dépend du choix des caractéristiques et de l'algorithme d'entraînement
- Cela amène aussi à examiner la question du partage des modèles appris sur des textes à caractère confidentiel.

Plan

Conclusion

- Besoin de désidentification des textes cliniques
- Bonnes performances après entraînement sur des textes annotés, sur le même type de texte
- Usage : Quel niveau de performance peut être considéré comme suffisant ?
- TAL : Autres pistes pour la création de corpus d'entraînement

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. De-identification of patient notes with recurrent neural networks. *J Am Med Inform Assoc*, Dec 30 2016.

Cyril Grouin and Pierre Zweigenbaum. Automatic de-identification of French clinical records: Comparison of rule-based and machine-learning approaches. In *Proc MEDINFO 2013*, Studies in Health Technology and Informatics, pages 476–480. Amsterdam, IOS Press, 2013. doi: doi:10.3233/978-1-61499-289-9-476.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270, San Diego, Ca., mai 2016. Association for Computational Linguistics.

Mohammed Saeed, Mauricio Villarroel, Andrew T. Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H. Kyaw, Benjamin Moody, and Roger G. Mark. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. *Critical Care Medicine*, 39:952–960, mai 2011.