Extraction d'information à partir de textes médicaux

Journées STP et Automatique de la SAGIP – 25-26 Novembre 2021

Angie NGUYEN, Doctorante

LAMIH CNRS, Arts et Métiers ParisTech, CEGEDIM R&D

angie.nguyen@ensam.eu

Sommaire

1

Introduction

Contexte et objectif

Introduction

- **95% des big data** sont des **données non structurées** (e.g. textes, audio, vidéo). Paradoxalement, la plupart des applications se sont focalisées sur les 5% de données structurées, qui ne représentent (Gandomi and Haider (2015)).
- Dans le secteur de la santé, une source de données riche mais peu exploitée sont les **documents médicaux** (e.g. comptes rendus d'hospitalisation, lettres au confrère).
- Deux facteurs inhibiteurs :
  - Difficulté d'exploiter les données (non structurées) en langage naturel.
  - Présence d'informations personnelles de santé.
- **Objectif : concevoir et développer un algorithme capable d'extraire de l'information médicale à partir de ces documents.**

2

Etat de l’art

Applications de l’extraction d’informations médicales

Dans la logistique pharmaceutique

- Dans un état de l'art systématique sur les *data analytics* en logistique pharmaceutique analysant 85 articles de 2012 à 2020, seulement deux contributions utilisent des données textuelles:
  - Balan (2018) ont développé un système d'extraction d'informations sur les bonnes pratiques environnementales en supply chain.
  - Tang (2019). ont publié un algorithme d'extraction de recommandations pour une meilleure gestion des stocks de produits pharmaceutiques.

- Analyse de sentiments des médias pour prédire la consommation de médicaments en période de crise (Nguyen 2021).

- Cependant, aucune contribution n'utilise des catégories d'informations médicales (e.g., symptômes, diagnostics).

Extraction d'information médicale : applications

- Un état de l'art publié en 2017 analyse **263 articles** appliquant l'extraction d'informations cliniques à partir de textes (wang et al.).

- **3 grands domaines** d'applications:
  - Études de maladie
  - Etudes de médicaments
    - Réactions aux médicaments
    - Toxicité
  - Optimisation des processus
    - Contrôle qualité
    - Traitement des patients

- Méthodes utilisées :
  - 65% approches par règles basées sur des thésaurus médicaux;
  - 35% approches par apprentissage automatique.
- SVM et régression logistique sont les modèles les plus utilisés.

Table 3
The most frequently used machine learning methods (top 6) and the corresponding number of papers in the included publications.

Method	No. of Papers
Support Vector Machine (SVM)	26
Logistic regression (LR)	11
Conditional random field (CRF)	9
Decision Tree (DT)	8
Naïve Bayes (NB)	6
Random Forest (RF)	4

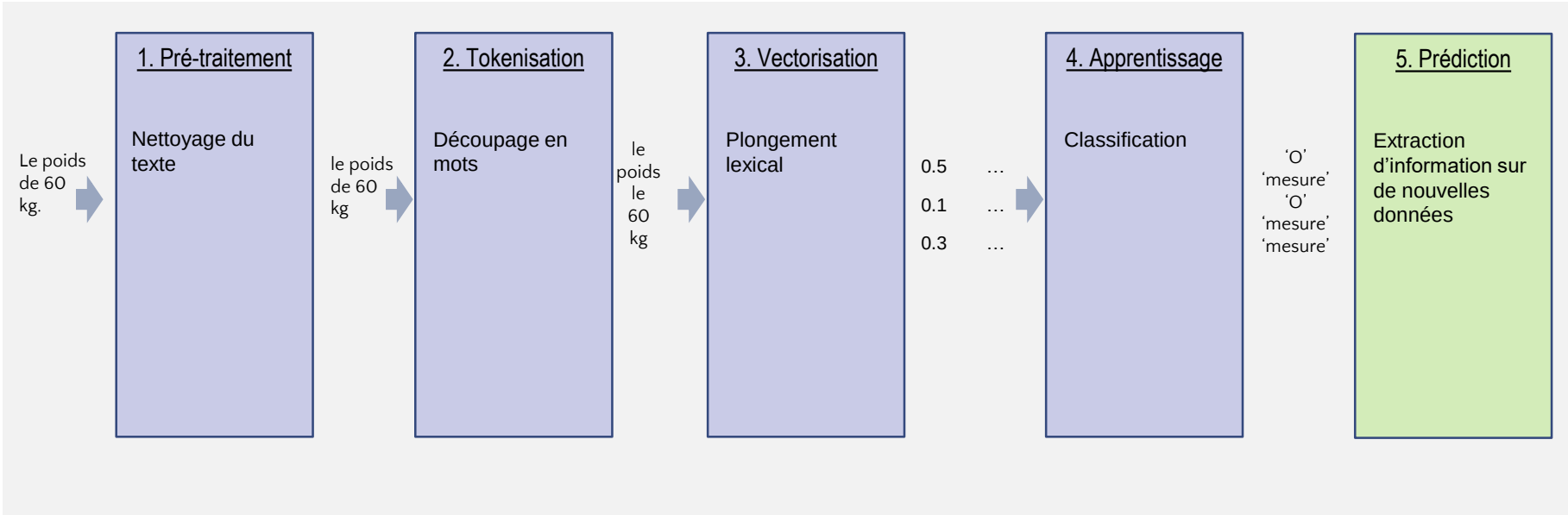
Source : wang et al. 2017

3

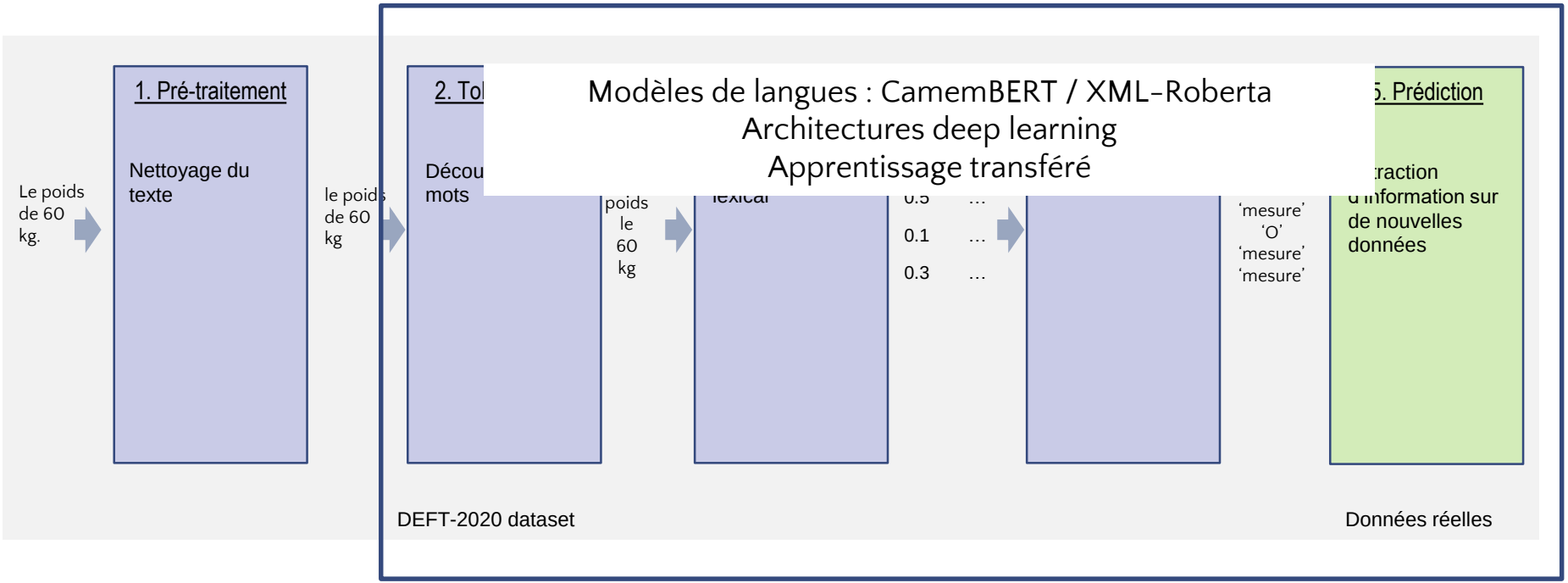
Méthodologie

Apprentissage transféré des modèles de langage

Méthodologie



Méthodologie



4

Résultats

Comparaison règles vs. apprentissage

Comparaison de l'approche par règles vs. par apprentissage

Par règles

Cher Confrère,
J'ai revu ce jour votre patient Mr suivi pour cancer du sein droit traité par mammectomie avec radiothérapie terminée en juin 2016.
Il va très bien sur le plan général et n'a pas de signe fonctionnel nouveau à signaler.
Le poids de 60 kg.
Cliniquement, je n'ai pas d'élément suspect à la palpation de la paroi droite, des aires ganglionnaires satellites.
Il m'apporte une échographie abdominale et une radio du thorax qui sont normales.
ACE et CA 15.3 sont normaux.
Il s'inquiète pour sa prostate et je lui demande donc un PSA.
A l'examen clinique, il n'y a pas d'anomalie à ce niveau.
Je le reverrai dans 6 mois, il doit poursuivre la prise de TAMOXIFENE.
En vous remerciant de votre confiance,
Je vous prie d'agréer, Cher Confrère, mes salutations dévouées.

diagnostic

spécialité

examen

médicament

Par apprentissage

Cher Confrère,
J'ai revu ce jour votre patient Mr suivi pour cancer du sein droit traité par mammectomie avec radiothérapie terminée en juin 2016.
Il va très bien sur le plan général et n'a pas de signe fonctionnel nouveau à signaler.
Le poids de 60 kg.
Cliniquement, je n'ai pas d'élément suspect à la palpation de la paroi droite, des aires ganglionnaires satellites.
Il m'apporte une échographie abdominale et une radio du thorax qui sont normales.
ACE et CA 15.3 sont normaux.
Il s'inquiète pour sa prostate et je lui demande donc un PSA.
A l'examen clinique, il n'y a pas d'anomalie à ce niveau.
Je le reverrai dans 6 mois, il doit poursuivre la prise de TAMOXIFENE.
En vous remerciant de votre confiance,
Je vous prie d'agréer, Cher Confrère, mes salutations dévouées.

Comparaison de l'approche par règles vs. par apprentissage

- Les méthodes basées sur les règles ont trois principaux avantages :
  - informations extraites structurées.
  - pas d'annotation de données.
  - explicabilité.

- Cependant, les méthodes basées sur l'apprentissage automatique sont plus performantes :
  - compréhension des équivalents, synonymes.
  - prise en compte du contexte (négations, conditions).

4 Perspectives

Perspectives

- Amélioration du modèle
  - Optimisation du modèle
  - Entraînement sur des données plus propres.
  - Enrichissement du modèle à l'aide de dictionnaires médicaux.
  - Approche hybride règle / apprentissage.

- Conséquences pour les industries en santé :
  - Enrichissement des bases de données de santé.

- Perspectives :
  - Surveillance épidémiologique (réseaux sociaux, médias).
  - Prédiction de la demande en produits.
  - Aide à la décision médicale par recherche intelligente (cas des maladies rares).

Références

Angie Nguyen, Samir Lamouri, Robert Pellerin, Simon Tamayo & Béranger Lekens (2021): Data analytics in pharmaceutical supply chains: state of the art, opportunities, and challenges, International Journal of Production Research, DOI: 10.1080/00207543.2021.1950937

Gandomi, Amir, and Murtaza Haider. 2015. “Beyond the Hype: Big Data Concepts, Methods, and Analytics.” International Journal of Information Management 35 (2): 137–144. doi:10.1016/j.ijinfomgt.2014.10.007.

Balan, Shilpa, and Sumali Conlon. 2018. “Text Analysis of Green Supply Chain Practices in Healthcare.” Journal of Computer Information Systems 58 (1): 30–38. doi:10.1080/08 874417.2016.1180654

Tang, Valerie, Paul K. Y. Siu, K. L. Choy, G. T. S. Ho, H. Y. Lam, and Y. P. Tsang. 2019. “A Web Mining-Based Case Adaptation Model for Quality Assurance of Pharmaceutical Warehouses.” International Journal of Logistics Research and Applications 22 (4): 325–348. doi:10.1080/13675567.2018. 1530204.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, Hongfang Liu, Clinical information extraction applications: A literature review, Journal of Biomedical Informatics, Volume 77, 2018.

Angie Nguyen, Samir Lamouri, Robert Pellerin, Managing demand volatility during unplanned events with sentiment analysis: a case study of the COVID-19 pandemic, IFAC-PapersOnLine, Volume 54, Issue 1, 2021.

Rémi Cardon, Natalia Grabar, Cyril Grouin, Thierry Hamon, Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques.
DEFT 2020, Atelier DÉfi Fouille de Textes, June 2020, Nancy, France. pp.1-13.

Cyril Grouin, Natalia Grabar, Vincent Claveau, Thierry Hamon, Clinical Case Reports for NLP, BIONLP 2019, 1st August 2019, Florence, Italy.

Extraction d'information à partir de textes médicaux

Journées STP et Automatique de la SAGIP – 25-26 Novembre 2021

Angie NGUYEN, Doctorante

LAMIH CNRS, Arts et Métiers ParisTech, CEGEDIM R&D

angie.nguyen@ensam.eu