Compréhension du langage naturel pour le dossier patient informatisé : accès à l'information et extraction d'information

Antoine Neuraz

Université de Paris

École doctorale Pierre Louis de Santé Publique (ED393)

Laboratoire INSERM UMRS 1138, Information Science for Precision Medicine

Compréhension du langage naturel pour le dossier patient informatisé

accès à l'information et extraction d'information

Par Antoine Neuraz

Thèse de doctorat d' Informatique médicale

Dirigée par Anita Burgun
Et par Sophie Rosset

Présentée et soutenue publiquement le 15 décembre 2020

Devant un jury composé de :

Marc CUGGIA, PU-PH, Université Rennes 1	Rapporteur
Benoit FAVRE, MCF, HDR, Aix-Marseille Université	Rapporteur
Guillaume ASSIE, PU-PH, Université de Paris	Examinateur
Christian LOVIS, PU, Université de Genève	Examinateur
Anita BURGUN, PU-PH, Université de Paris	Directrice de thèse
Sophie ROSSET, DR, Université Paris-Saclay, CNRS	Codirectrice de thèse

Remerciements

Je tiens à remercier Anita Burgun, ma directrice, pour son soutien sans faille et ses conseils avisés, ainsi que ma co-directrice, Sophie Rosset, qui a su croire en moi malgré de longues périodes où j'ai pu être considéré comme perdu de vue.

Je remercie chaleureusement Marc Cuggia et Benoit Favre qui ont accepté d'être rapporteurs de ce travail. Je tiens également à remercier les autres membres du jury, Guillaume Assié et Christian Lovis, d'avoir accepté de participer. Il a été étonnamment facile de trouver une date qui convienne à tous, merci pour votre disponibilité.

Je voudrais aussi remercier mes acolytes : Nicolas Garcelon, Bastien Rance et Maxime Wack, piliers de… l'informatique médicale internationale, du monde entier et de l'univers.

Je remercie également la team HEGP, William Digan, Rosy Tsopra, Alice Rogier, David Baudoin pour leur aide précieuse.

Je remercie Ivan Lerner, Jordan Jouffroy et Sarah Feldman, qui ont fait, ~~parfois~~ chacun à leur façon, avancer le projet.

Je souhaite également remercier Leonardo Campillos Llanos, Sahar Ghannay, Pierre Zweigenbaum, Aurélie Névéol, Cyril Grouin et tous les membres du groupe ILES du LIMSI pour leur expertise et nombre de discussions très intéressantes.

Bien entendu, je remercie aussi tous les membres de l'Équipe 22, future équipe HeKA (on y croit), parce qu'*Aucun de nous ne sait ce que nous savons tous, ensemble* [Euripide]. (Oui j'ai regardé sur internet)

Je remercie mes proches pour leur soutien et leurs encouragements.

Je remercie également tous les lecteurs de ces pages, qui, avisés, ne dépasserons probablement pas cette page de remerciements.

Table des matières

Abréviations 1

Chapitre 1 : Introduction **3**
1.1 Traitement automatique de la langue médicale 4
1.2 Les modèles d'apprentissage utilisés en traitement automatique de la langue . . . 6
1.3 Interroger le dossier patient en langue naturelle 15
1.4 Restructurer l'information présente sous forme de texte libre dans le dossier patient 20

Chapitre 2 : Compréhension de la langue naturelle pour le dialogue orienté tâche dans le domaine biomédical dans un contexte de faibles ressources **23**
2.1 Introduction 23
2.2 Méthode 24
2.3 Génération et augmentation de données 24
2.4 Expériences 30
2.5 Résultats 31
2.6 Discussion 34
2.7 Conclusion 36
2.8 *Article : Natural language understanding for task oriented dialog in the biomedical domain in a low resources context* 38
2.9 *Article : The Impact of Specialized Corpora for Word Embeddings in Natural Langage Understanding* 49

Chapitre 3 : APMed : un corpus de textes cliniques en français annotés pour les informations sur les médicaments **55**
3.1 Matériels et méthodes 55
3.2 Résultats 61
3.3 Discussion 63
3.4 *Article : MedExt : combining expert knowledge and deep learning for medication extraction from French clinical texts* 66

Chapitre 4 : PyMedExt : une trousse à outils pour l'annotation de texte et l'échange de données dans le domaine clinique **85**
4.1 Introduction 85
4.2 Les composants de PyMedext 86
4.3 Conclusion 90

Chapitre 5 : Traitement de la langue naturelle pour une réponse rapide aux maladies émergentes : application au traitement par inhibiteurs calciques pendant la pandémie de COVID-19 . . . . . **91**

5.1 Materiels et méthodes . . . 91
5.2 Resultats . . . 96
5.3 Discussion . . . 100

Chapitre 6 : Conclusion . . . 111

Annexe A : Guide d'annotation pour les données médicamenteuses dans les textes cliniques en français . . . 115
A.1 Overview : . . . 115
A.2 Medication : . . . 116
A.3 Dosage : . . . 119
A.4 Frequency . . . 120
A.5 Duration . . . 122
A.6 Mode of administration . . . 123
A.7 Condition . . . 124
A.8 8.Events . . . 125
A.9 Attributes . . . 127
A.10 Prescription Pattern . . . 128

Annexe B : Table des questions "vie réelle" . . . 131

Annexe C : Questions "vie réelles" en représentation hierarchique . . . 139

References . . . 147

Liste des tableaux

2.1 Description des mentions 27
2.2 Charactéristiques des jeux de données 30
2.3 Résultat des expériences sur la tâche de labellisation de séquence 31
2.4 Résultats des expériences sur la tâche de classification 32
2.5 Comparaison des différents types d'embedding 34

3.1 Description des entités annotées 57
3.2 Description des événements affectant les entités annotées 57
3.3 Description des attributs modifiant les entités annotées 57
3.4 Description des annotations dans le corpus APMed 61
3.5 Résultats moyens des modèles BiLSTM-CRF sur le corpus APMed 62
3.6 Résultat du système d'annotation hybride sur le corpus APMed 63

5.1 Noms et codes ATC des inhibiteurs calciques 95
5.2 Noms et codes CIM10 des phénotypes 95
5.3 Description des données disponibles en fonction des sources 96
5.4 Performances du modèle d'extraction d'informations sur les médicaments AVANT normalisation des entités. 97
5.5 Performances du modèle d'extraction d'informations sur les médicaments APRES normalisation des entités. 97
5.6 Description de la population de l'étude. 99
5.7 Modèle de Cox multivarié (CI : Interval de confiance, HR hazard ratio). 100

B.1 Questions 'vie réelle' 132

C.1 Questions 'vie réelle' 140

Table des figures

1.1 Architecture d'un neurone artificiel (source : Wikipedia) 7
1.2 Convolution 1D avec noyau de taille 3 et pas de 1 (source : peltarion.com) 8
1.3 1D max pooling avec une fenêtre de taille 2 et pas de 2 (source : peltarion.com) . 8
1.4 Réseau de neurones récurrents (source : wikipedia.org) 9
1.5 Long short-term memory (source : wikipedia.org) 9
1.6 Gated Recurent Unit (source : wikipedia.org) 10
1.7 RNN et RNN bidirectionnel (source : wikipedia.org) 11
1.8 encodeur-décodeur (source : smerity.com) 11
1.9 encodeur-décodeur avec attention (source : Bahdanau,2015) 12
1.10 Architecture du transformer (source : Vaswani, 2017) 14
1.11 Répresentation du vocabulaire : one-hot versus word embeddings 14
1.12 Panorama des agents conversationels en santé. Adapté de Montenegro *et al.*[144]. 17
1.13 Architecture d'un agent conversationnel 18
1.14 Exemple du déroulé d'un échange avec un agent conversationnel 19

2.1 Shcéma expérimentatl 25
2.2 Modélisation des questions relatives aux examens de biologie. en bleu, les modalités de questions envisagées a priori. en blanc, les autres modalités découvertes dans les questions réelles. 26
2.3 Combinaison des modèles et modifieurs pour le jeu de données d'entrainement. . 28
2.4 Labellisation de séquence 32
2.5 Classification 33
2.6 Exemple de parsing utilisant une approche hiérarchique combinant intention et entités 35

3.1 Deux exemples d'annotations 56
3.2 Schéma du modèle hybride d'annotation 58
3.3 Schéma du modèle d'annotation basé sur BERT 60

4.1 Vue générale de PyMedExt 86
4.2 Les différentes classes de PyMedExt 87
4.3 La classe Document de PyMedExt 88

5.1 Description du pipeline de traitement automatique du language 92
5.2 Exemples d'expressions régulières pour l'extraction de phénotypes 94
5.3 Comparaison des données en fonction de leur provenance 96
5.4 Flowchart 98
5.5 Temporalité des données en fonction de leur provenance 99

À Céline, Sarah, Annah et Raphaël
parce que c'est un travail d'équipe

Abréviations

Sigle	Définition
APHP	Assistance-Publique Hôpitaux de Paris
ASGD	asynchronous stochastic gradient descent
ATC	Anatomical Therapeutic Chemical
ATIS	Air Travel Information System
BERT	Bidirectional Encoder Representations for Transformers
BiLSTM	Bidirectional LSTM
CIM10	classification internationale des maladies version 10
CNN	Convolutional neural network, réseaux de neurones convolutionnels
CR	Compte-rendu
CRF	Conditional random fields
CRH	Compte-rendu d'hospitalisation
DPI	Dossiers patients informatisés
EDS	Entrepôt de données de santé l'APHP
EHR	Electronic health record (=DPI)
FHIR	Fast Healthcare Interoperability Resources
GRU	gated recurrent units
HL7	Health Level Seven International
HMM	modèles de Markovs cachés
IC	inhibiteurs calciques
IM	Information médicale
JMIR	Journal of Medical Internet Research
LSTM	Long short term memory unit
NER	extraction d'entités nommées
NLP	Natural language processing (=TAL)
OMOP CMD	modèle de données commun de l'Observational Medical Outcomes Partnership
POS	Part-of-speech (POS)
PPDB	Paraphrase Database
RELU	rectified linear unit
RNN	Recurrent neural network, réseaux de neurones récurrents
RNNG	Recurrent neural network grammars
SVM	support vector machines
TAL	traitement automatique du langage
TALm	traitement automatique du langage médical
UIMA	Unstructured Information Management applications
UMLS	Unified Medical Language System

Chapitre 1

Introduction

Le dossier patient informatisé (DPI) pose encore aujourd'hui des problèmes d'utilisabilité et de maintenance [1,2]. L'evidence based medicine (la médecine basée sur les preuves) a incité depuis plusieurs années les hôpitaux à structurer l'acquisition des données à travers des formulaires plus ou moins standardisés dans le DPI. Les données structurées sont généralement associées à un référentiel et stockées sous forme tabulaire. Un exemple de données structurées est le codage des diagnostics selon une terminologie standardisée (*e.g.*, classification internationale des maladies version 10, CIM10) stockés dans une table. L'avantage de ce type de données est qu'elles sont faciles à requêter, à comptabiliser, à partager et à analyser. Elles ont cependant des inconvénients importants. Leur expressivité, c'est-à-dire la variété d'information qu'elles peuvent véhiculer, est limitée *a priori* par les référentiels utilisés. Une pathologie absente du référentiel ne pourra pas être codée et donc l'information sera perdue. Ce type de situation peut survenir quand les connaissances évoluent plus rapidement que les référentiels comme c'est le cas pour les maladies génétiques rares ou les maladies émergentes. L'utilisabilité des données structurées dépend de la complexité du référentiel utilisé. De ce fait, si l'utilisateur ne maîtrise pas parfaitement le référentiel, il ne pourra pas coder l'information la plus adéquate. Les référentiels utilisent parfois une terminologie qui diffère de l'usage des utilisateurs ce qui peut rendre délicate leur utilisation en pratique quotidienne. Il est alors fréquent que les cases *commentaires* des questionnaires de recueil d'informations structurées soient remplies d'informations pertinentes que l'utilisateur n'aura pas su renseigner autrement. Enfin, le recueil d'informations structurées incite à ne renseigner que des informations avec un degré de certitude élevé. Quand un seul champ est disponible et que plusieurs hypothèses sont valides, soit une seule des hypothèses sera renseignée, soit aucune. Toute la subtilité et la richesse de la réflexion clinique sont alors perdues. Là encore, les maladies rares ou émergentes requièrent le recueil le plus fin possible, sous une forme narrative de l'histoire de la maladie, des signes, symptômes et tableaux cliniques.

Ces éléments expliquent pourquoi, malgré toutes les avancées techniques récentes pour structurer le DPI, la médecine moderne s'appuie encore beaucoup sur le texte libre et la langue naturelle pour conserver et échanger des informations [3]. Les médecins écrivent ou dictent des lettres et des comptes-rendus pour adresser un patient à un confrère ou pour garder une trace d'une consultation ou d'un acte qu'ils ont effectué. Les examens d'imagerie ou d'anatomopathologie sont accompagnés de comptes-rendus dans lesquels les spécialistes précisent la description des images et leur interprétation.

Outre les questions relatives à la documentation et la réutilisation des données cliniques, se pose la question de l'utilisabilité des systèmes d'information tels qu'ils existent aujourd'hui.[1,2]

Une revue systématique de 2005 pointait que le temps de documentation avait tendance à être augmenté avec l'utilisation des DPI contrairement à la promesse initiale [4] Dans la même étude, l'utilisation de logiciels de prescription pouvait faire augmenter le temps nécessaire à cette tâche de 98.1% à 328.6%. Pire, Gardner *et al* en 2019, dans une étude sur 1800 médecins dans la région de Rhodes Island aux États-Unis, ont mis en évidence une association entre l'utilisation de DPI et un risque accru de burnout [5]. Dans cette étude, 70% des médecins montraientt des signes de stress liés à l'utilisation des outils informatiques médicaux.

Il y a toujours une tension entre les besoins d'expressivité, de communication et de fluidité qui sont favorisées par l'utilisation fde la langue naturelle et les besoins de réutilisation ultérieure des informations qui ont souvent amèné à réclamer des données plus structurées.[6] Cependant, nous pouvons faire l'hypothèse que maximiser les possibilités apportées par la langue naturelle, aussi bien pour la recherche d'information que pour la documentation permettra de réduire ces tensions en réduisant la quantité de données structurées à compléter tout en maintenant une expressivité maximale. Il s'agit alors de développer des méthodes de restructuration du texte afin d'optimiser sa réutilisabilité.

Dans cette optique, nous proposons de nous intéresser à deux types d'utilisation de la langue naturelle en relation avec le dossier patient :

1. pour formuler une requête, afin de répondre à une question comme : *Le patient a-t-il eu une anémie depuis sa dernière opération ?* ;
2. pour restructurer de l'information en langue naturelle présente dans les textes cliniques comme par exemple l'extraction de l'entité CIM10 "I21 - Infarctus aigu du myocarde" dans la phrase : *Patient de 38 ans, adressé pour une suspicion d'infarctus du myocarde.*

Nous allons dans un premier temps introduire le champ du traitement automatique de la langue médical et décrire un certain nombre de méthodes utiles pour la suite de l'exposé. Nous reviendrons ensuite sur les deux axes que nous venons d'esquisser qui sont l'interrogation du dossier patient en langue naturelle et la restructuration des informations de dossier patient présentes dans le texte libre.

1.1 Traitement automatique de la langue médicale

Le traitement automatique de la langue dans le domaine médical (TALm) existe depuis plusieurs décennies. Cependant, l'augmentation de la quantité de données disponibles et l'augmentation de la puissance des ordinateurs ont fait émerger de nouvelles tendances. En effet différentes revues de la littérature sur le TALm mettent en évidence un glissement au fur et à mesure des années de systèmes purement basés sur des règles expertes, vers des méthodes probabilistiques puis, plus récemment vers des méthodes d'apprentissage automatique (machine learning).[7–9] Le principe étant de faire apprendre à l'ordinateur à réaliser une tâche donnée en entraînant un algorithme à réaliser cette tâche à partir d'une série d'exemples pour lesquels la réponse est connue. Ces méthodes d'apprentissage sont dites supervisées. La supervision venant du guidage de l'algorithme avec des exemples pré-résolus pendant la phase d'apprentissage. Ce qui s'oppose aux méthodes d'apprentissage non supervisées (ou auto-supervisées) pour lesquelles seules les données d'entrées sont nécessaires. Concrètement, l'approche supervisée repose sur une série d'étapes clés :

- Définition précise de la tâche : *Extraction des noms de médicaments dans les comptes-rendus cliniques*
- Élaboration d'un guide d'annotation précisant quels éléments doivent être annotés et la façon de les annoter : *le nom de médicament comprend-il le dosage ? Si oui dans quelles conditions ? les alimentations parentérales sont-elles considérées comme des médicaments ?*
- Constitution d'un jeu de données : *Extraction de X textes de comptes-rendus cliniques dans l'entrepôt de données de l'hôpital Y*
- Annotation du jeu de données par un ou plusieurs annotateurs (gold-standard). Ce jeu de données annotées servira en partie à l'apprentissage et en partie à l'évaluation.
- Préparation des données. *prétraitement du texte, transformation des mots en valeurs numériques, enrichissement avec calcul d'attributs spécifiques, séparation des jeux de données d'entraînement, et de test*
- Entraînement d'un modèle d'apprentissage à partir d'une partie du jeu de données annotées (jeu d'entraînement). Le modèle s'entraîne en essayant de minimiser l'erreur de prédiction sur le jeu de données d'entraînement. La façon de mesurer l'erreur est définie en fonction de la tâche et en fonction de critères d'évaluation définis a priori.
- Évaluation des performances du modèle. L'évaluation des performances se fait à partir d'une autre partie du jeu de données annotées (jeu de test) gardées de côté pour l'occasion. Le modèle entraîné effectue la tâche sur le jeu de données test et l'on compare les résultats avec les données du gold-standard.

Il existe un certain nombre de tâches de TALm qui incluent les tâches génériques de traitement du texte comme la segmentation (découpage des textes en sections, paragraphes, phrases), la tokenisation en mots (découpage des mots au sein de la phrase) qui produit des tokens, l'étiquetage morpho-syntaxique (Part-of-speech tagging, associe à chaque token ses attributs grammaticaux), la lemmatisation (associe à chaque token son lemme, sa forme canonique), ou encore l'analyse syntaxique (défini un arbre syntaxique qui représente la structure de relations entre les tokens d'une phrase basée sur leur fonction syntaxique). Mais la richesse du TALm se situe plutôt au niveau sémantique. En effet, un grand nombres de tâches sémantiques se retrouvent dans le TALm :

La reconnaissance d'entités nommées consiste à segmenter une séquence de mots pour en extraire un groupe correspondant à une entité présente dans un référentiel. Les personnes, les lieux ou les organisations sont des entités nommées classiques. Dans le domaine médical, les entités nommées pourront être des maladies, des phénotypes, des molécules, des gènes, des localisations anatomiques, etc.

La détection d'attributs consiste à labelliser des mots, ou groupes de mots avec des éléments qui modifient leur sens. Par exemple, la détection de la polarité (e.g. négation, affirmation), du degré de factualité (e.g., avéré, hypothèse, conditionnel), du sujet (concerne le patient, sa famille, etc), de la temporalité

La normalisation des entités consiste à retrouver le lien entre la mention d'une entité dans une phrase (e.g. *infarctus du myocarde*, *IDM*) avec le concept correspondant dans une terminologie de référence (I21 : Infarctus aigu du myocarde).

L'extraction de relations consiste à identifier les éléments qui sont en relation et à catégoriser cette relation. Dans la phrase, *La pneumopathie bactérienne est une infection.* Les entités *pneumopathie bactérienne* et *infection* sont reliées par une relation de type hiérarchique *is-a*.

La traduction consiste à traduire une séquence de tokens depuis une source vers une

représentation cible. Cette représentation cible pouvant être une autre langue, mais également une représentation formelle.

La résolution de coréférences consiste à identifier différents groupes de tokens (le plus souvent des syntagmes nominaux) faisant référence dans le discours à la même entité. Dans, *Le patient était fébrile. Il a été traité avec du paracétamol.*, *Le patient* et *Il* se réfèrent à la même personne.

L'extraction de la temporalité consiste à identifier les relations temporelles entre les différents éléments extraits d'un texte.

1.2 Les modèles d'apprentissage utilisés en traitement automatique de la langue

1.2.1 Les champs aléatoires conditionnels

Les champs aléatoires conditionnels (conditional random fields, CRF) sont utilisés pour des tâches de labellisation de séquence. Ils se retrouvent ainsi en bioinformatique, en extraction d'information ou en reconnaissance de la parole).[10]

Les CRF sont dérivés des modèles de Markovs cachés (HMM) qui sont des automates probabilistes à état fini. Les HMM sont des modèles génératifs qui cherchent à définir la probabilité jointe $P(X, Y)$ avec X la séquence d'observation et Y la séquence d'étiquettes. Ils font l'hypothèse d'une indépendance des observations afin de contourner l'explosion combinatoire. Les CRF quant à eux adoptent une approche conditionnelle, c'est à dire qu'ils visent à déterminer la probabilité conditionnelle $P(Y|X)$, la probabilité des séquences d'étiquettes pour une séquence d'observation donnée. En traitement de la langue, les CRF les plus communs sont les CRF en chaîne linéaire de premier ordre.[11]

1.2.2 Réseaux de neurones artificiels

Neurone artificiel

L'histoire des réseaux de neurones artificiels n'est pas récente puisqu'on en retrouve les prémices en 1943.[12] Le perceptron fut lui décrit par Rosenblatt en 1957 et popularisé par Minsky et Papert en 1969.[13] Un neurone artificiel est une structure mathématique qui mappe une série d'entrées x_i avec i avec une sortie y.(Figure 1.1)

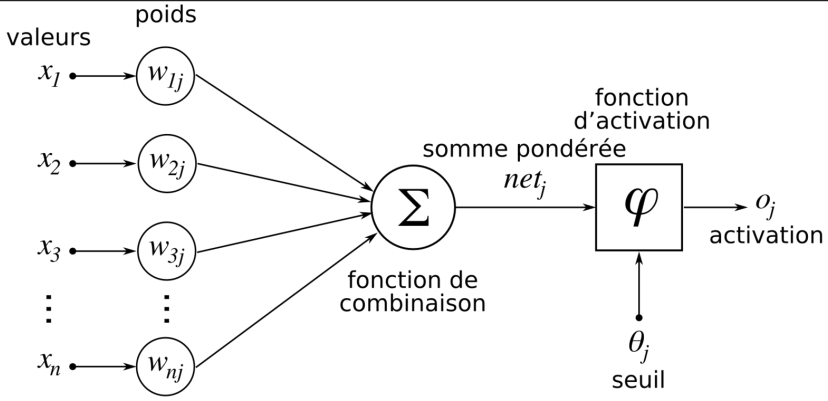


FIGURE 1.1 : Architecture d'un neurone artificiel (source : Wikipedia)

Pour un neurone k, avec m différentes entrées x_i et $m+1$ poids w_i (un poids supplémentaire correspondant au biais), la sortie est définie par :

$$y_k = \varphi(\sum_{j=0}^{m} w_{jk} x_j)$$

avec φ étant une fonction d'activation, typiquement une fonction comme tanh ou rectified linear unit (RELU) [14].

Back-propagation

L'algorithme de back-propagation (rétropropagation du gradient) qui permit l'entraînement de réseaux multicouches apparut lui en 1986.[15] Ce mécanisme d'optimisation consiste à corriger les erreurs en fonction de l'importance des éléments qui participent de cette erreur.[16] La back-propagation utilise la descente de gradient pour minimiser la loss L. Ainsi ce sont les poids qui contribuent à une erreur importante qui sont le plus modifiés. L'erreur se propage $e_i^{(n)} \mapsto e_j^{(n-1)}$ comme suit :

$$e_j^{(n-1)} = \varphi'^{(n-1)}(h_j^{n-1}) \sum_i w_{ij}^{(n)} e_i^{(n)}$$

avec $e_i^{(n)} = e_i^{sortie} = (y_i - t_i)\frac{\delta y_i}{\delta h_i^n}$; φ' étant la dérivée de la fonction d'activation φ, t_i la véritable sortie attendue, et h la fonction d'agrégation (souvent la somme pondérée des poids et des entrées du neurone)

Réseaux de neurones convolutionels

Les réseaux convolutionels (convolutional neural network, CNN) sont un type de réseau de neurones acycliques (feed-forward) qui mime le cortex visuel des animaux. Ils permettent, en vision par ordinateur de repérer des petits motifs patterns invariants indépendamment de leur localisation dans l'image.[17] Il s'agit alors de convolution en 2 dimensions. Dans le cadre du texte, le mécanisme est assez similaire, les CNN vont permettre de repérer des groupes de tokens, des n-grams de tokens. La détection de ces groupes de tokens ne dépendant pas de leur position dans la séquence. Ainsi, la séquence globale n'est pas prise en compte. La taille des n-grams détectée est définie par la taille du noyau de convolution l.(Figure 1.2)

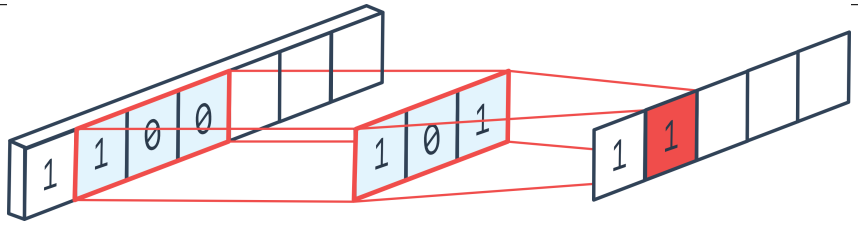


FIGURE 1.2 : Convolution 1D avec noyau de taille 3 et pas de 1 (source : peltarion.com)

Si l'on prend la séquence de tokens $x_1...x_n$, pour un noyeau de taille l, si l'on se place à la position i :

$$conv(x, i) = \varphi(W * concat([x_{i-l/2} : x_{i+l/2}]) + b)$$

avec $concat()$ qui réalise la concaténation des vecteurs $[x_{i-l/2} : x_{i+l/2}]$ et les paramètres entrainables W et b. Le filtre de convolution se déplace le long de la séquence par pas de s (avec $s = 1$ dans l'illustration). Il est possible de combiner plusieurs couches de convolutions en parallèle ou en séquentiel avec des tailles de noyau différentes afin de capturer différentes informations.

A la suite d'une étape de convolution, il est d'usage d'ajouter une une couche de "pooling" qui va combiner les information de plusieurs positions. En général, on utilise une fonction maximum, max-pooling, pour ne garder que la valeur maximale d'activation dans la fenêtre de pooling. (Figure 1.3) Le max-pooling permet de renforcer le signal des groupes importants.

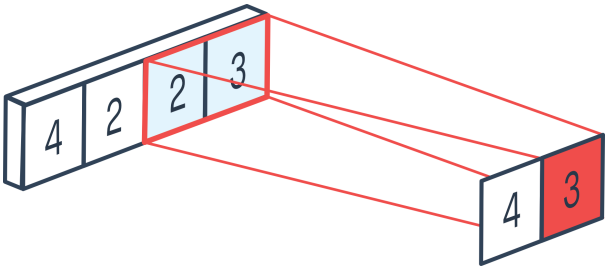


FIGURE 1.3 : 1D max pooling avec une fenêtre de taille 2 et pas de 2 (source : peltarion.com)

RNN

Les réseaux de neurones récurrents (recurrent neural networks, RNN) sont très utilisés en TAL car ils sont particulièrement adaptés à l'analyse des séquences. Les neurones d'un RNN sont connectées par des connexions récurrentes. Ces connexions possèdent un poids qui a la particularité d'être partagé entre les neurones. L'entrée d'un neurone est caractérisée par la combinaison d'un élément de la séquence au temps t x_t avec la sortie du neurone précédent h_{t-1}. Le réseau est paramétrisé par les matrices W, U et b. (1.4)

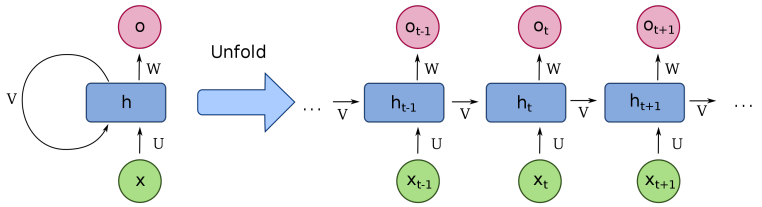


FIGURE 1.4 : Réseau de neurones récurrents (source : wikipedia.org)

L'état h_t au temps t ainsi que la sortie finale du réseau sont donnés par les équations suivantes :

$$h_t = \varphi_h(W_r x_t + U_r h_{t-1} + b_h)$$
$$y_t = \varphi_h(W_y h_t + b_y)$$

avec : x_t le vecteur d'entrée, h_t le vecteur de la couche cachée (état latent), y_t le vecteur de sortie, W, U et b les paramètres, φ_h et φ_y les fonctions d'activations.

Lors de l'entraînement, la retro-propagation du gradient se fait le long de la séquence en commençant par la fin. La correction propagée diminue exponentiellement avec la taille de la séquence. Ainsi, lorsque la séquence s'alonge, le gradient a tendance à disparaître. Pour palier à cette difficulté, d'autres types de cellules de réseaux récurrents ont été proposés : les Long short-term memory (LSTM) et les gated recurrent units (GRU)

Long Short-Term Memory Les LSTM ont été introduit par Hochreiter *et al* en 1997.[18] L'idée a été d'une part d'ajouter une mémoire d'état c_t en plus de l'état caché h_t, d'autre part 3 portes qui contrôlent la mise à jour (input gate) I_t, la sortie (output gate) O_t et l'oubli de l'état antérieur (forget gate) F_t. La transmission de c_t se fait à gain constant et évite la disparition du gradient. (Figure 1.5)

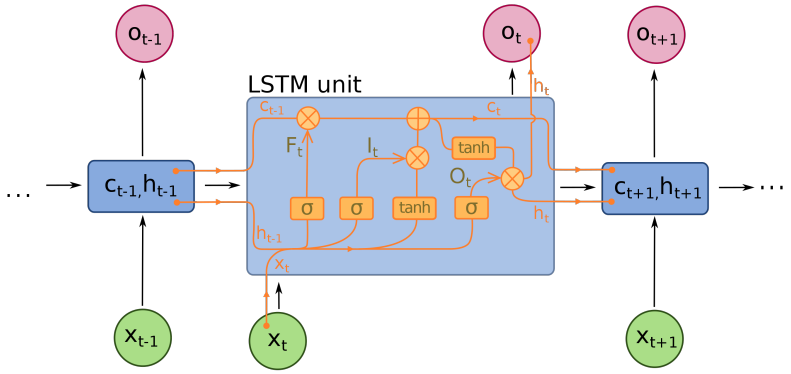


FIGURE 1.5 : Long short-term memory (source : wikipedia.org)

L'état est initialisé avec les valeurs $c_0 = 0$ et $h_0 = 0$

$$
\begin{aligned}
F_t &= \sigma(W_F x_t + U_F h_{t-1} + b_F) && \text{(forget gate)} \\
I_t &= \sigma(W_I x_t + U_I h_{t-1} + b_I) && \text{(input gate)} \\
O_t &= \sigma(W_O x_t + U_O h_{t-1} + b_O) && \text{(output gate)} \\
c_t &= F_t \circ c_{t-1} + I_t \circ \tanh(W_c x_t + U_c h_{t-1} + b_c) \\
h_t &= O_t \circ \tanh(c_t) \\
o_t &= \varphi(W_O h_t + b_o)
\end{aligned}
$$

Gated Recurent Unit Les GRU ont été introduites par Chung et al en 2014. [19] Il s'agit d'une version simplifiée des LSTM mais qui permet de maintenir de bonnes performances. Les input gate et forget gate sont fusionnées en une update gate Z_t, la output gate est remplacée par une reset gate R_t et la mémoire c_t disparaît. Comme il y a moins de paramètres, l'apprentissage est plus rapide. (Figure 1.6)

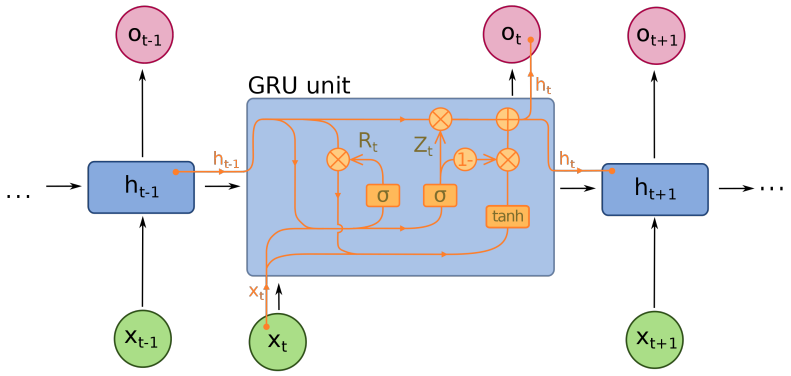


FIGURE 1.6 : Gated Recurent Unit (source : wikipedia.org)

$$
\begin{aligned}
Z_t &= \sigma(W_Z x_t + U_Z h_{t-1} + b_Z) && \text{(update gate)} \\
R_t &= \sigma(W_R x_t + U_R h_{t-1} + b_R) && \text{(update gate)} \\
h_t &= Z_t \circ h_{t-1} + (1 - Z_t) \circ \tanh(W_h x_t + U_h(R_t \circ h_{t-1} + b_h))
\end{aligned}
$$

Réseaux bidirectionnels Le principe des réseaux bidirectionnels (biRNN) est d'utiliser deux couches de RNN différentes qui parcourent la séquence en sens inverse puis de concaténer les sorties. L'idée est de permettre à chaque temps de bénéficier de la représentation latente des contextes provenant de l'amont et de l'aval de la séquence. Ils ont été inventés par Schuster et Paliwal en 1997.[20] (Figure 1.7)

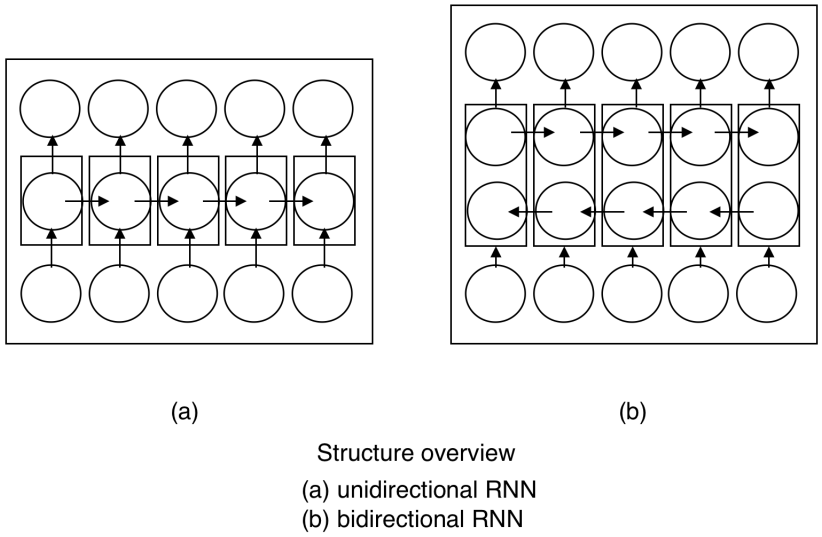


FIGURE 1.7 : RNN et RNN bidirectionnel (source : wikipedia.org)

En traitement de la langue, ils sont utilisés pour de la traduction [21], de l'analyse de dépendance[22], ou encore de la reconnaissance d'entités nommées [23].

Les modèles seq2seq (encodeur-decodeur) Les modèles de type encodeur-décodeur (sequence to sequence, seq2seq) consistent à utiliser deux RNN distincts. Le premier va parcourir l'ensemble de la séquence d'entrée (éventuellement avec un biRNN) et renvoyer le dernier état latent qui correspond à une représentation de l'ensemble de la séquence. Cette représentation est ensuite fournie en entrée à un deuxième RNN qui va produire une séquence de sortie (ou une sortie unique). Ces modèles sont particulièrement bien adaptés aux tâches de traduction car ils permettent de découpler l'ordre et la taille de la séquence de sortie par rapport à la séquence d'entrée.[24] (Figure 1.8)

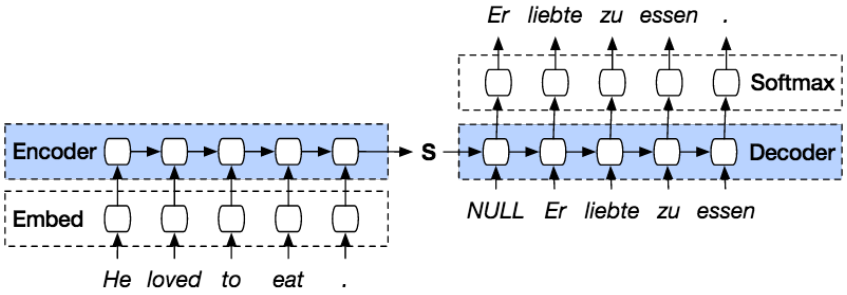


FIGURE 1.8 : encodeur-décodeur (source : smerity.com)

Les mécanismes d'attention Les mécanismes d'attention sont apparus pour compléter les seq2seq sur des tâches de traductions avec de longues séquences. Le problème du seq2seq est qu'il n'utilise en entrée du décodeur que le dernier état latent. Lorsque la séquence est longue, il est possible que ce seul état ne puisse pas véhiculer suffisamment d'information. L'idée princeps de l'attention était de permettre au décodeur d'accéder indirectement aux états de chaque élément de l'encodeur tout en pondérant ces informations.[24] (Figure 1.9)

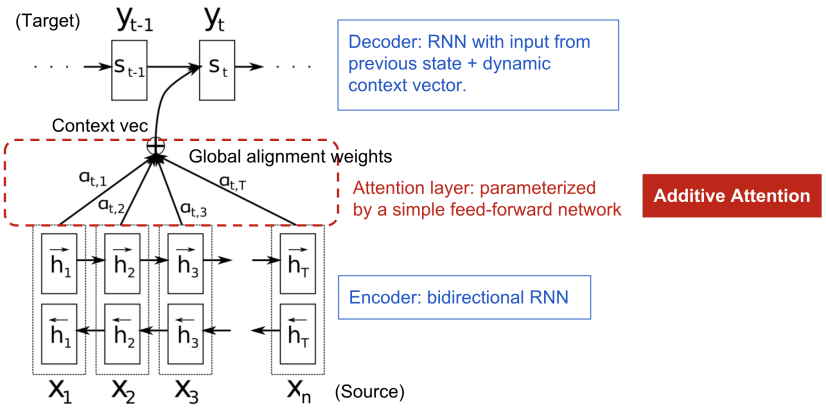


FIGURE 1.9 : encodeur-décodeur avec attention (source : Bahdanau,2015)

Ce mécanisme permet un alignement entre la source et la cible contrôlé par un vecteur de contexte c_t. Le vecteur de contexte est produit à partir des états latents de l'encodeur h_t, des états latents du décodeur s_t et de pondérations d'alignement $\alpha_{t,i}$.

$$\alpha_{t,i} = \frac{\exp(\text{score}(s_{t-1}, h_i))}{\sum_{i'=1}^{n} \exp(\text{score}(s_{t-1} h_{i'}))} \quad \text{softmax d'un score prédéfini}$$

$$c_t = \sum_{i=1}^{n} \alpha_{t,i} \quad \text{vecteur de contexte de } y_t$$

Le vecteur d'alignement α est paramétrisé à l'aide d'un simple réseau feed-forward avec une seule couche qui est entraîné conjointement avec les autres parties du réseau. Bahdanau et al [24] définissaient la fonction de score comme suit :

$$\text{score}(s_t, h_i) = v_a^T \tanh(W_a[s_t; h_i])$$

avec v_a et W_a étant des matrices de paramètres à optimiser. Il existe différents types de mécanismes d'attention en fonction de la façon de définir la fonction de score : content-base attention [25], additive [24], location-base [26], dot-product [26], scaled dot-product [27].

Nous pouvons aussi distinguer l'attention locale (porte sur une partie de l'espace d'entrée) ou globale (porte sur la totalité de l'espace d'entrée) [26,28] ou encore la self-attention (met en relation différentes positions de la même séquence d'entrée) [29].

Le Transformer

Le transformer est une architecture proposée par Vaswani *et al* en 2017 [27]. Il s'agit d'un modèle qui mime les possibilités offertes par un seq2seq mais sans réseau récurrent. L'architecture est entièrement basée sur des mécanisme de self-attention. (Figure 1.10)

Multi-head self-attention L'unité de base est une multi-head self-attention. Cette unité prend trois entrées : Clé (K), Valeur (V), Requête (Q). Les représentations encodées des entrées sont vues comme des paires (K, V) de dimension n (la taille de la séquence). En fonction des tâches, les clés et les valeurs peuvent être identiques. Dans le décodeur, la sortie précédente est

compressée en une requête Q de dimension m. La sortie suivante est produite à partir de ces trois composants.

L'opération de base est la scaled-dot product attention qui est une somme pondérée des valeurs dans laquelle le poids assigné à chaque valeur provient du dot-product de la requête avec les clés :

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QT^T}{\sqrt{(}n)})V$$

On parle de multi-head car l'unité repose sur un mécanisme d'ensembling qui calcule en parallèle un nombre h d'attentions au lieu d'une seule. Les résultats sont ensuite concaténés et transformés linéairement pour obtenir les dimensions adéquates :

$$\begin{aligned}\text{MultHead}(Q, K, V) &= [head_1; ...; head_h]W^O \\ \text{avec head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)\end{aligned}$$

où W_i^Q, W_i^K, W_i^V et W_i^O sont des matrices de paramètres.

Encodeur Chaque couche d'encodage est composée d'une multi-head attention unit suivie d'une couche de normalisation suivie d'une simple couche dense feed-forward suivie d'une nouvelle couche de normalisation. Chacune de ces 2 sous-couches sont connectées par des connexions résiduelles.

Décodeur Chaque couche de décodeur est composée de deux sous-couches de multi-head attention et d'une sous-unité feed-forward, chacune étant suivie d'une couche de normalisation et pourvue de connexions résiduelles. La première couche de multi-head attention est modifiée pour éviter que l'attention ne puisse accéder aux positions qui suivent celle qui est en cours de prédiction.

Architecture complète D'abord la source et la cible sont encodées à l'aide d'embeddings et produisent des données de même dimension. Pour préserver l'information sur la position dans la séquence, les embeddings sont modifiés en ajoutant le résultat d'une fonction sinusoïde (positional encoding). Enfin, une tête est ajoutée à la suite de l'encodeur comprenant une couche linéaire dense et une softmax.

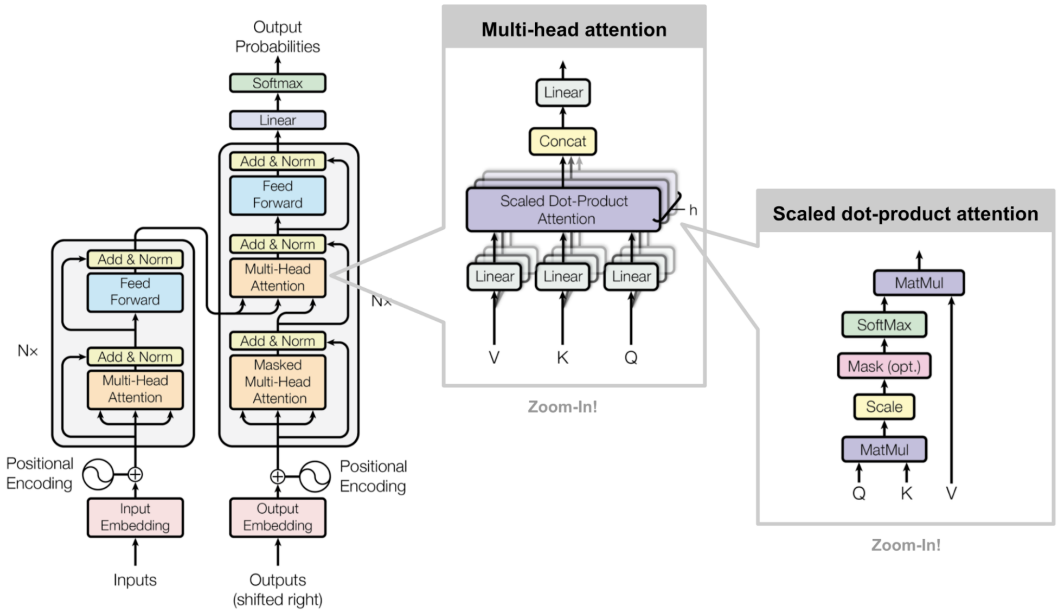


FIGURE 1.10 : Architecture du transformer (source : Vaswani, 2017)

Les embeddings

Un des éléments qui a le plus fait progresser les performances des algorithmes de deep-learning sur les tâches de TAL durant les dernières années a été l'apparition des plongements lexicaux (word-embeddings). Traditionnellement en apprentissage, le texte était représenté en utilisant un encodage de type one-hot. C'est à dire qu'on utilise un vecteur de la taille du vocabulaire dont tout les éléments sont à 0 sauf élément correspondant au token présent. Cela conduit à une représentation très sparse, c'est à dire dans laquelle l'essentiel des éléments sont à 0. D'autre part, cette représentation ne contient pas d'information sur la distance sémantique codée dans le modèle. Le principe des word embeddings est de construire une représentation dense du vocabulaire avec une dimension très inférieure à la taille du vocabulaire et qui intègre des informations sur la distance sémantique entre les mots. (Figure 1.11)

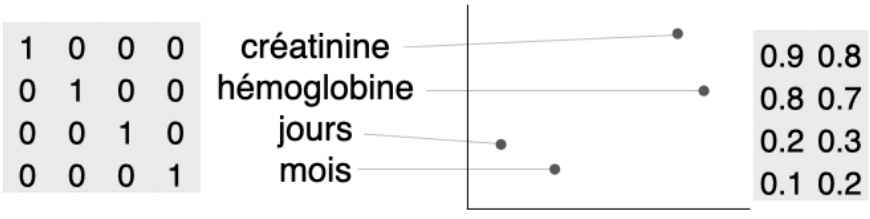


FIGURE 1.11 : Répresentation du vocabulaire : one-hot versus word embeddings

Plusieurs méthodes ont été proposées mais elles ont toutes en commun l'avantage d'être auto-supervisées. C'est à dire qu'elles peuvent être entrainées sur des données textuelles brutes, non annotées. Il en découle que l'on peut donc utiliser de très larges corpora pour entraîner ces représentations, des corpora plus grands et plus riches que ceux annotés spécifiquement

pour une tâche. Il est donc possible pour les tâches de prédiction ultérieures de bénéficier de représentations riches pour des mots qui sont absents dans le jeu de données d'entraînement de la tâche (mais qui sont présents dans le large dataset qui a servi à entraîner l'embedding).

La méthode pionnière en la matière est l'algorithme continuous skip-gram de word2vec [30] qui utilise une fenêtre de contexte autour du mot pour calculer sa représentation. Il s'agit d'un simple réseau de neurones qui est chargé de prédire les mots de la fenêtre de contexte à partir du mot. A la même période, le modèle Glove calculait ses représentations à partir d'une matrice de co-occurrences. [31]. En se basant sur l'algorithme continuous skip-gram, fastText proposait d'intégrer des représentations de ngrams de caractères afin de pallier l'absence de représentation des mots hors vocabulaire.[32]

La principale limite de ces méthodes est qu'elles fournissent un unique vecteur par mot quel que soit son contexte d'utilisation. De nouvelles méthodes ont été proposées pour pallier cette limite : les embeddings contextuels. ELMo par exemple, utilise la représentation latente d'un modèle de langue bidirectionnel (se basant sur un biLSTM) pour proposer une représentation contextualisée.[33] De plus, ELMo utilise un embedding de caractère suivi d'un CNN en amont du biLSTM de façon à pouvoir proposer des représentations des mots inconnus (mots hors vocabulaire, out of vocabulary words, OOV). Plus récemment encore, le modèle BERT, pour Bidirectional Encoder Representations from Transformers, a provoqué un petit séisme dans la communauté TAL en produisant des performances au delà de l'état de l'art sur de nombreuses tâches.[34] BERT se base sur l'architecture transformer que nous avons décrite plus tôt. Les transformers sont intrinsèquement bidirectionnels puisque la séquence est lue dans son intégralité en une seule passe. Les représentations qui en découlent peuvent donc tirer pleinement partie de l'ensemble du contexte. BERT est entraîné sur une tâche de masked language modeling : le modèle est entraîné à prédire des mots masqués dans une phrase. Ces modèles sont très performants mais ils ont un très grand nombre de paramètres à optimiser et nécessitent de grandes quantités de données et d'importantes ressources pour l'entraînement. Le principe est donc de n'entraîner ces modèles depuis le départ qu'une seule fois et, quand une nouvelle tâche se présente, de ne faire qu'un fine-tuning (ne réentraîner qu'une partie des paramètres, par exemple les dernières couches en fixant les premières.) Cette étape de fine-tuning étant beaucoup moins gourmande en données et en ressources informatiques. De nombreuses variations autour de ces architectures ont été proposées depuis, proposant des variations d'architecture et/ou de méthodes d'entraînement. Nous pouvons citer de manière non exhaustive : ELECTRA[35], ALBERT [36], RoBERTa[37]. Certains modèles ont été spécifiquement entraînés sur des corpora en français : FlauBERT[38], CamemBERT[39].

1.3 Interroger le dossier patient en langue naturelle

La première tâche, interroger le dossier patient, s'entend dans le cadre d'une interface conversationnelle. L'objectif est de rendre le requêtage des dossiers plus efficace en permettant au médecin d'aller au delà de la recherche par mot clé ou même la recherche structurée. S'il s'agit de retrouver une information simple, les recherches par mots clé peuvent être très efficaces et sont souvent implémentées dans les DPI. Mais dès lors que le nombre de critères augmente ou que la requête implique des informations multimodales (provenant de plusieurs sources), les interfaces traditionnelles sont vites dépassées et le temps augmente rapidement.[1] La langue naturelle, de part sa compositionalité et sa compacité permet d'exprimer des requêtes parfois complexes de

manière efficace. Cela peut réduire également la charge cognitive imposée à l'utilisateur puisque comme son nom l'indique, la langue naturelle est la façon dont on a l'habitude naturellement de s'exprimer.

ELIZA en 1966 [40], fut le premier chatbot de ce type, chargé de simuler une conversation avec un psychothérapeute. Depuis, la littérature a prouv" l'intérêt potentiel de l'utilisation d'agents conversationnels dans le cadre de la santé. Des essais cliniques randomisés ont même montré le bénéfice de ces outils sur l'activité physique, l'accessibilité à l'information de santé en ligne ou la consommation de fruits et de légumes.[41–43] Cependant ces systèmes sont limités car les possibilités d'expression des utilisateurs sont contraintes, par exemple, par des questions à choix multiples. Depuis quelques années, l'augmentation des performances des systèmes d'intelligence artificielle a relancé l'intérêt pour ce type d'interfaces. Dans le domaine général, le développement exponentiel des assistants intelligent produit par des grandes firmes de technologies en sont la preuve. En effet, les avancées en machine learning et particulièrement concernant les réseaux de neurones, ont permis le développement d'outils plus complexes en termes de gestion du dialogue et plus souples pour l'utilisateur.[44]

Il y a un intérêt croissant pour la recherche sur les interfaces conversationnelles dans le domaine médical.[45] (Figure 1.12) Il existe des systèmes de dialogues orienté objectif (*goal oriented*) qui sont conçus pour répondre à des objectifs précis (e.g., donner des informations sur une maladie ou un traitement). Il existe également des systèmes de dialogues ouverts, dont l'objectif est de maintenir une conversation sans but particulier (chatbots). Des systèmes de dialogue dans le domaine de la santé ont été proposés dans diverses applications pour assister les patients au quotidien [46–52], pour aider à l'entraînement des patients (troubles autistiques [53], réminiscence [54], comportement [55]), ou des médecins et étudiants en médecine [56–60], aider au diagnostic [61–63], aider à l'éducation des patients [64–67] ou encore à la prévention [68–70]. A destination du médecin également, pour aider à la rédaction de rapports [71,72], pour retrouver des images [73] ; aider à la décision [74,75], aider au recueil d'information [76–78].

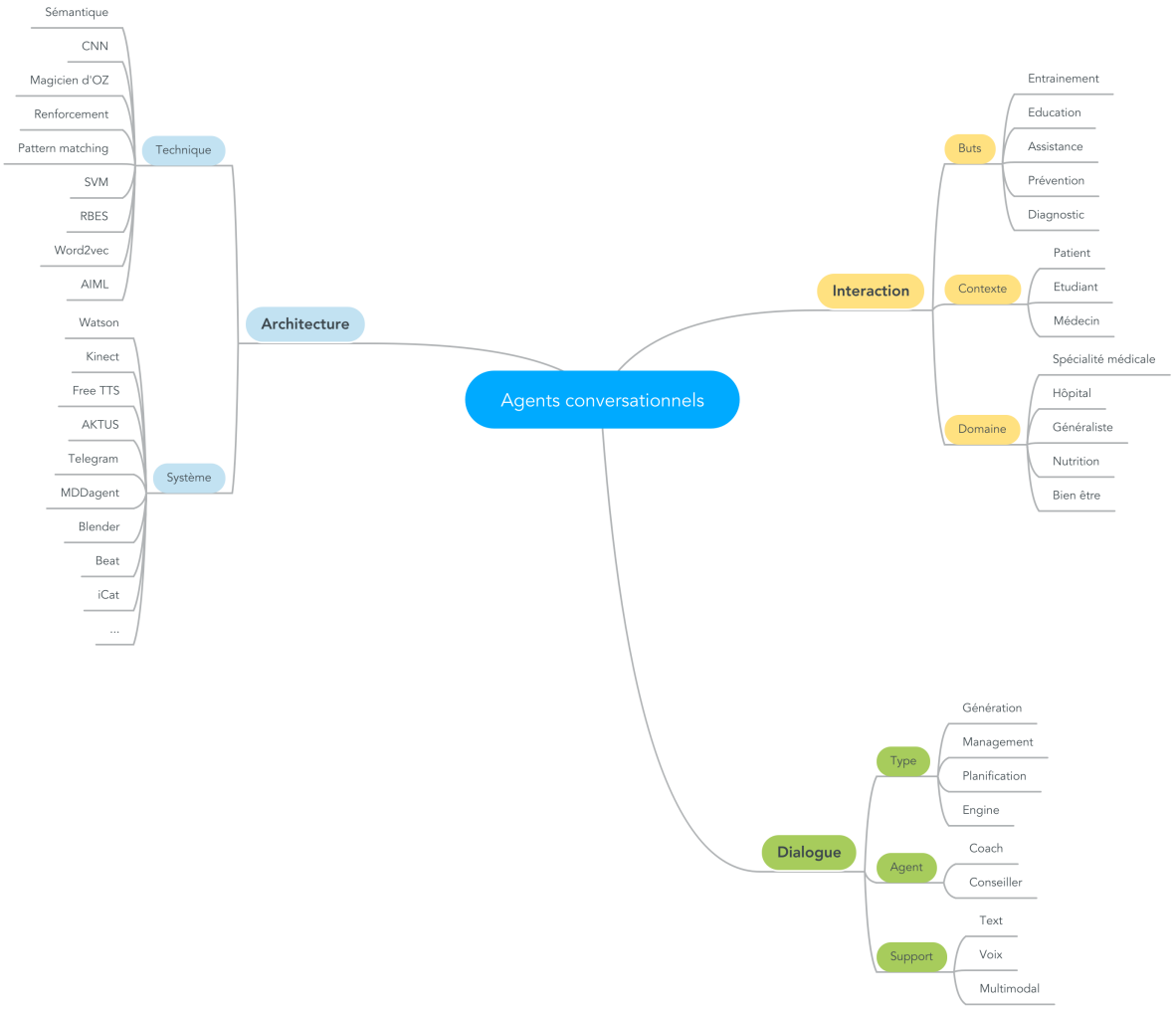


FIGURE 1.12 : Panorama des agents conversationels en santé. Adapté de Montenegro *et al.*[144].

Les systèmes de dialogue sont constitués de plusieurs composants [79] : un module de compréhension de la langue (*natural langage understanding*, NLU) qui permet de formaliser la langue naturelle, un module de gestion du dialogue qui est chargé de maintenir le suivi de l'état du dialogue et gérer les actes de dialogues, un module de génération de texte afin de produire les réponses du système et éventuellement, un module de requête vers une base de données et un module qui permet de modéliser les connaissances.(Figures 1.13 et 1.14)

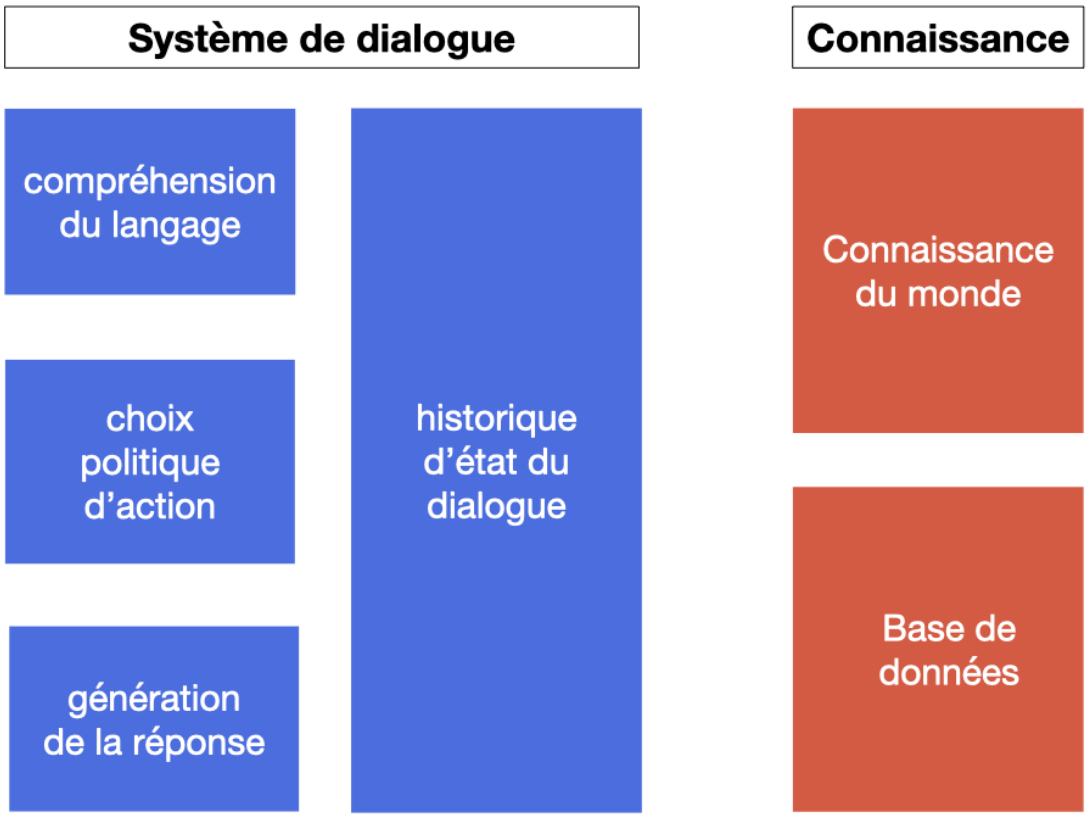


FIGURE 1.13 : Architecture d'un agent conversationnel

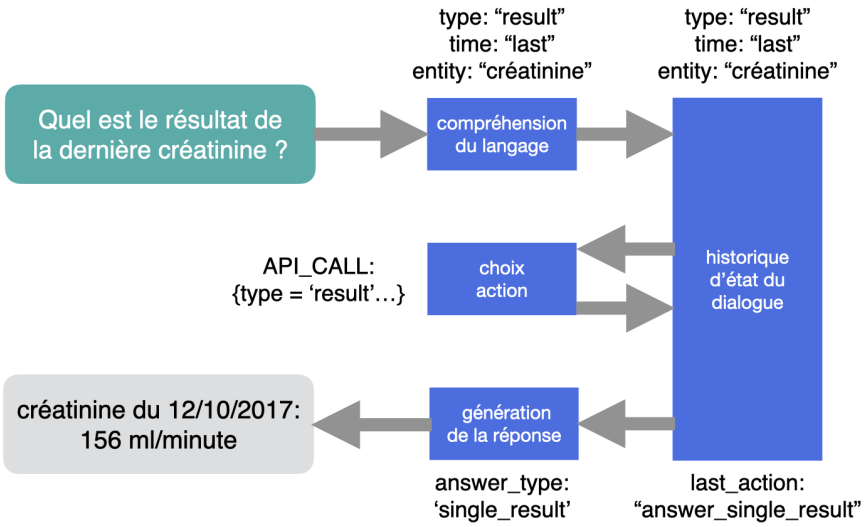


FIGURE 1.14 : Exemple du déroulé d’un échange avec un agent conversationnel

Notre recherche bibliographique ne nous a pas permis d’identifier de système qui permette d’interroger le dossier patient en langue naturelle, encore moins en français.

Nous nous plaçons donc dans la perspective du développement d’un agent conversationnel qui permette d’interroger le dossier patient en langue naturelle et nous intéressont au premier élément nécessaire à un tel outil, c’est à dire le module de compréhension de la langue.

La compréhension de la langue dans le cadre d’un agent conversationnel consiste en deux sous tâches : d’une part une labellisation de la séquence de mots pour repérer les entités d’intérêt (slot-filling) et d’autre part, l’identification de l’intention de l’utilisateur qui est une tâche de classification de texte.

Cette spécialité du TAL a suivi la même trajectoire historique que le reste de la discipline. Les premiers systèmes étaient basés sur des règles expertes et des méthodes à base de dictionnaires.[@ 40] Un autre courant, basé essentiellement sur les caractéristiques linguistiques s’est attaché à construire des capacités de compréhension génériques en visant à produire des formes formelles de langage.[80] Ces deux approches ont continué à coexister au cours du temps et ont été complétées à partir des années 90 par les systèmes statistiques. Le développement de jeux de données spécialisés et annotés ont permis l’expansion de ces approches dans des secteurs autres que la médecine (e.g., Air Travel Information System (ATIS) [81], MEDIA [82]). Les approches statistiques les plus utilisées furent les HMM [83], les méthodes de classification discriminatives (e.g. arbres de décisions)[84], les approches à base de connaissances et les grammaires hors contexte (context-free grammars)[85,86] et enfin les CRF [87,88]. Nous référons le lecteur à une revue très complète de ces méthodes par Tur et Mori.[89]

Les dernières années ont vu l’arrivée des approches basées sur les réseaux de neurones profonds. Les premières utilisations de ce type de méthodes pour la détection de l’intention (classification) remontent à 2006 avec les deep belief networks de Hinton *et al.*[90] Aujourd’hui, les approches les

plus communes sont à base de CNN [91–93], de RNN [94,95], la combinaisons de CNN et de RNN [96], des CNN avec mécanisme d'attention [97], ou des réseaux hierarchiques avec attention[98].

Pour le slot-filling, même si les CNN sont utilisés [99], les architectures les plus communes sont des variantes de RNN[100] : LSTM [101], RNN-EM (RNN avec mémoire externe) [102], LSTM encodeur-décodeur [103], seq2seq combiné avec un pointer network.[104]

Bien entendu toutes ces méthodes ne peuvent s'envisager qu'en la présence de jeux de données d'entraînement disponibles. Or, nous avons pu constater qu'aucun système de ce type n'avait été proposé pour le DPI. Il n'y a donc pas de jeu de données disponible. Une méthode classique pour pallier ce déficit est de développer un système complet à base de règles et de commencer à recueillir des données utilisateur afin de construire un jeu de données d'entraînement. Cette approche a l'inconvénient d'imposer le développement d'un système complet, suffisamment robuste pour ne pas faire fuir les utilisateurs immédiatement. D'autre part, les systèmes à base de règle sont complexes, aussi bien à développer qu'à maintenir.

Nous avons donc choisi d'adopter une approche différente et hybride : générer un jeu de données à partir de quelques exemples experts et l'augmenter par différentes méthodes afin d'entraîner un système de NLU. Ce système pouvant ensuite être confronté aux utilisateurs pour enrichir le set d'exemples voire réentraîner un modèle à partir du nouveau set de données acquis. Pour la génération du jeu de données, nous avons tiré parti des terminologies existantes, utilisé des modèles de questions, évalué la génération de paraphrases par traduction automatique et étudié l'apport d'embeddings contextuels (ELMo) entraînés sur différents datasets (domaine général ou spécialisé). Cette étude est détaillée dans le Chapitre 2.

1.4 Restructurer l'information présente sous forme de texte libre dans le dossier patient

Si l'on s'intéresse à l'extraction d'information médicale à partir des textes de dossiers patients, il existe une littérature riche et nous pouvons citer ici quelques revues pertinentes [8,9,105–107].

Nous nous sommes ici focalisés sur les méthodes d'extraction concernant les données sur les médicaments et leurs attributs, mais les méthodes et tendances sont transposables à d'autres types de donnés comme les pathologies par exemple.[108] Pourquoi nous intéresser au médicament en particulier ? Principalement parce que c'est un problème majeur de santé publique. En 2018, en France, les dépenses concernant les médicaments en ville étaient de 32.7 milliards d'euros, représentant 16% du budget de la santé.[109] Les effets indésirables liés aux médicaments sont un important problème de santé publique puisqu'ils sont fatals pour 0.15% des patients. De plus, un tiers des hospitalisations dues à des effets indésirables sont évitables, souvent associées à un historique médicamenteux mal renseigné ou des effets indésirables rares.[110,111] Une partie non négligeable de cet historique médicamenteux existe le plus souvent sous forme non structurée dans les textes rédigés par les médecins ou les équipes soignantes lors de la prise en charge des patients.[112] Ces textes sont regroupés, dans les DPI. Cependant, leur forme non structurée les rend difficilement utilisables en pratique, dès lors que la complexité du dossier augmente.[105] Un certain nombre de travaux adressent ce problème spécifique depuis plusieurs années pour extraire le nom des médicaments, le dosage, la fréquence de prise, la durée, la voie d'administration, la raison et les conditions de prise, ou encore détecter les effets indésirables.

1.4.1 Méthodes d'extraction des informations sur les médicaments

Les premières approches étaient plutôt basées sur des méthodes de règles expertes associées à l'utilisation de dictionnaires terminologiques et de variantes lexicales, avec des systèmes comme MedEx ou MedLEE.[113–118]. Des systèmes basés sur l'apprentissage automatisé ont également été proposés. Avant l'arrivée des réseaux de neurones profonds, les principaux algorithmes utilisés pour cette tâche étaient les *support vector machines* (SVM) et les CRF.[119–122]

Les travaux les plus récents se basent essentiellement sur des approches à base de réseaux de neurones profonds. Plus précisément, pour la reconnaissance d'entités nommées qui nous intéresse ici, les réseaux de neurones récurrents sont les plus utilisés.[123] Ces architectures, spécialisées dans le traitement des séquences sont particulièrement bien adaptées pour traiter le texte puisqu'elles permettent de calculer une représentation de l'ensemble de la séquence (dans notre cas, de la phrase). Actuellement, l'état de l'art est constitué par des algorithmes de type biLSTM-CRF[124] qui combinent des unités biLSTM avec des CRF.[125–130] Bien entendu, les SVMs et CRFs sont toujours utilisés par certaines équipes avec des résultats parfois comparables aux approches par RNN.[131,132] Même si les BiLSTM-CRF sont les plus populaires, d'autres variantes de réseaux de neurones ont également été proposées. Comme par exemple, l'ajout d'un mécanisme d'attention au LSTM[133], la combinaison de CNN et de LSTM [134], ou encore l'utilisation combinée d'ensembles de modèles.[135–137]

1.4.2 Jeux de données d'entraînement

La grande majorité des systèmes présentés dans la section précédente sont destinés à traiter des textes en anglais. D'une manière générale, l'anglais est prédominant dans le domaine du TAL et plus particulièrement dans le TAL médical.[138] Cela s'explique d'une part par la surreprésentation de l'anglais sur internet en général et dans la littérature médicale en particulier ; ainsi que par la domination des jeux de données d'entraînement en anglais proposés pour des challenges de TAL. Les challenges de TAL permettent aux équipes de comparer les modèles qu'ils développent sur des données communes afin d'assurer la comparabilité des résultats. Ces jeux de données sont annotés manuellement et constituent donc des *gold standard* permettant d'entraîner des modèles (train set) et de les évaluer (test set). De manière facultative, un jeu de développement (*dev set*) qui permet d'ajuster les hyper-paramètres des modèles sans dévoiler le *test set*. Dans le domaine de l'extraction de données sur les médicaments, un jeu de données, réalisé en 2009 a permis de grandes avancées : le challenge i2b2 2009 medication. Il contenait 1243 résumés de sortie (696 pour le *train set* et 547 pour le *test set*) provenant de Partners Healthcare.[139] Ce jeu de données a servi de base au développement au travail présenté ici. Par la suite, deux autres jeux de données, ont été réalisés : le challenge n2c2 2018 et MADE 1.0.[140,141] Le challenge n2c2 2018, issu de la base MIMIC-III [142] inclue 505 résumés de sortie (303 pour l'entraînement, 202 pour le test), et contient 9 types de concepts cliniques (e.g. nom du médicament) et 8 attributs (e.g. fréquence, durée, voie d'administration, effet secondaire) ainsi que les relations entre les 8 attributs et les médicaments.

Du fait de l'absence de jeux de données d'entraînement disponibles pour le français, très peu de systèmes ont été proposés pour cette langue. En réalité, nous n'avons trouvé qu'une seule référence sur l'extraction d'informations médicamenteuses sur les textes médicaux en français par Deléger et al. en 2010.[143] Il s'agit d'un système à base de règles.

Nous avons donc choisi de développer un jeu de données comprenant des textes cliniques en français issus de dossiers patients informatisés, annoté pour les informations sur les médicaments. A partir de ce dataset nous avons développé des modèles d'extraction d'information médicamenteuses utilisant des embeddings contextuels (ELMo, BERT), des RNN de type BiLSTM-CRF et des approches hybrides (règles expertes) et séquentielles. Le développement de ce dataset et de ces modèles est décrit dans le Chapitre 3.

Afin de répondre aux nécessités relatives à la mise en production de modèles de ce type au sein d'un système d'information hospitalier, nous avons également développé PyMedExt, une trousse à outils pour l'annotation de texte et l'échange de données dans le domaine clinique. Cet outil permet de faciliter importation et l'exportation de données depuis et vers des formats d'échange de données cliniques ainsi que la conversion à partir de formats d'annotations TAL. Ce travail est décrit au Chapitre 4.

Enfin, nous nous montrons comment ces modèles d'extraction et ces outils peuvent être utilisés pour aider à répondre à des questions de santé publique. Durant la première vague de COVID-19, nous avons déployé notre pipelines d'extraction d'information clinique sur l'entrepôt de données cliniques de l'APHP. Nous avons ensuite réalisé une étude d'épidémiologie clinique qui met en évidence l'apport des données extraites par ce pipeline. Cette étude est décrite au Chapitre 5

Chapitre 2

Compréhension de la langue naturelle pour le dialogue orienté tâche dans le domaine biomédical dans un contexte de faibles ressources

2.1 Introduction

Accéder à l'information pertinente de manière rapide et fiable est un élément essentiel pour la pratique médicale quotidienne. Quand un médecin recherche une information, il est important qu'il puisse l'obtenir immédiatement ou avec un délai très faible. Le temps passé à rechercher l'information représente au minimum un manque d'efficience, au pire une perte de chances pour le patient.

Or, l'information médicale (IM) est complexe. L'IM est très hétérogène, présente à la fois sous forme structurée et non structurée (e.g., texte, valeurs numériques, codages, images), et provenant de sources différentes, humaines ou automatiques et de lieux différents, même au sein d'un même hôpital (e.g., services cliniques, laboratoires, services d'imagerie). De plus, l'IM est produite par des professionnels à destination d'autres professionnels. De ce fait, elle contient une part importante d'élision, et de faits qui doivent être interprétés à l'aune de connaissances supposées acquises par le lecteur expert. Pour ajouter à la complexité, l'IM est très relationnelle : différentes informations, stockées à des endroits différents peuvent être reliées par de nombreux types de relations (e.g., hiérarchie, causalité, synonymie). Par exemple, si un médecin veut savoir si son patient à une anémie - définie par un diminution du taux d'hémoglobine dans le sang - il pourra aller chercher l'information sur le taux d'hémoglobine dans les résultats d'examens biologiques, mais également dans un compte-rendu médical qui pourra mentionner l'anémie du patient ou encore dans des codes diagnostics de visites précédentes. Enfin, l'IM est temporelle : elle est datée et change au cours du temps. La dynamique d'un paramètre est parfois plus importante que sa valeur absolue. Le médecin pourra donc être amené à rechercher l'évolution d'un paramètre, son augmentation ou sa diminution, sa fréquence de survenue, ou la date de dernière occurrence.

Du fait de cette complexité, les interfaces classiques de recherche d'information peinent souvent

à répondre aux attentes des médecins. Pouvoir interroger le dossier patient en langue naturelle permettrait une expressivité et une fluidité beaucoup plus importantes.

Comme nous avons pu le voir dans le Chapitre 1, il y a un intérêt croissant pour la recherche sur les interfaces conversationnelles dans le domaine médical.[45]. Notre recherche bibliographique ne nous a pas permis de trouver de système de dialogue adressant spécifiquement la recherche d'information dans le dossier d'un patient, encore moins en français.

Dans ce travail, nous nous sommes intéressé au premier module d'un agent conversationnel de ce type, celui de compréhension de la langue. Pour développer un tel module en utilisant une approche basée sur l'apprentissage automatisé, un des premiers prérequis serait un jeu de données d'entraînement. Ce jeu de données requiert des énoncés d'utilisateurs - ce sont les entrées du système de NLU - associées à leur représentation formelle - ce qui constitue la sortie du système de NLU. De plus, il doit être suffisamment grand et diversifié pour être représentatif de la tâche. Mais comment faire si un tel jeu de données n'existe pas ? En fonction de la tâche et de la langue, il est probable qu'on ne puisse pas trouver de jeu de données satisfaisant. Dans ces circonstances, nous pouvons dire que les ressources sont faibles en terme de données. En effet, un régime de faible ressources peut apparaître dans différentes conditions. La grande majorité des études en NLP sont faites sur l'anglais. Ainsi, la grande majorité des outils et des jeux de données sont faits pour l'anglais. Même dans des langues bien pourvues, si la tâche est très spécifique, dans un domaine très spécialisé, il est probable de ne pas pouvoir trouver de données suffisantes. Dans le cadre des approches par apprentissage statistique, qui reposent sur de grands jeux de données d'entraînement, un régime de faibles ressources contraindra méthodes possibles et limitera les performances des modèles appris. Le domaine médical est un bon cas d'usage en ce qui concerne les faibles ressources en terme de données, particulièrement pour des langages autres que l'anglais. En effet, du fait des restrictions liées à la confidentialité des données, il est difficile de partager de vraies données médicales issues de dossiers patients.

Une solution pour circonvenir à l'absence de jeu de données pour entraîner un modèle est de générer un jeu de données basé sur un faible nombre d'exemples et d'utiliser des techniques de supervision distante pour augmenter la taille et la couverture du jeu de données via l'utilisation de terminologies, intégration de connaissances externes et la génération de paraphrases. Dans ce travail, nous avons étudié comment des modèles, entraînés avec un tel jeu de données en français, se comportaient face à un jeu de données réel.

2.2 Méthode

2.3 Génération et augmentation de données

Dans cette section, nous décrivons la tache et les méthodes utilisées pour générer les données d'entraînement et de développement. Nous expliquons les méthodes de génération de données à partir de modèles (*templates*) et de terminologies, la génération de paraphrases de ces modèles par traduction en pivot et l'incorporation de connaissances externes via des word embeddings (plongements lexicaux en français) et des modèles de langage. La figure 2.1 détaille le schéma général des expériences de ce travail.

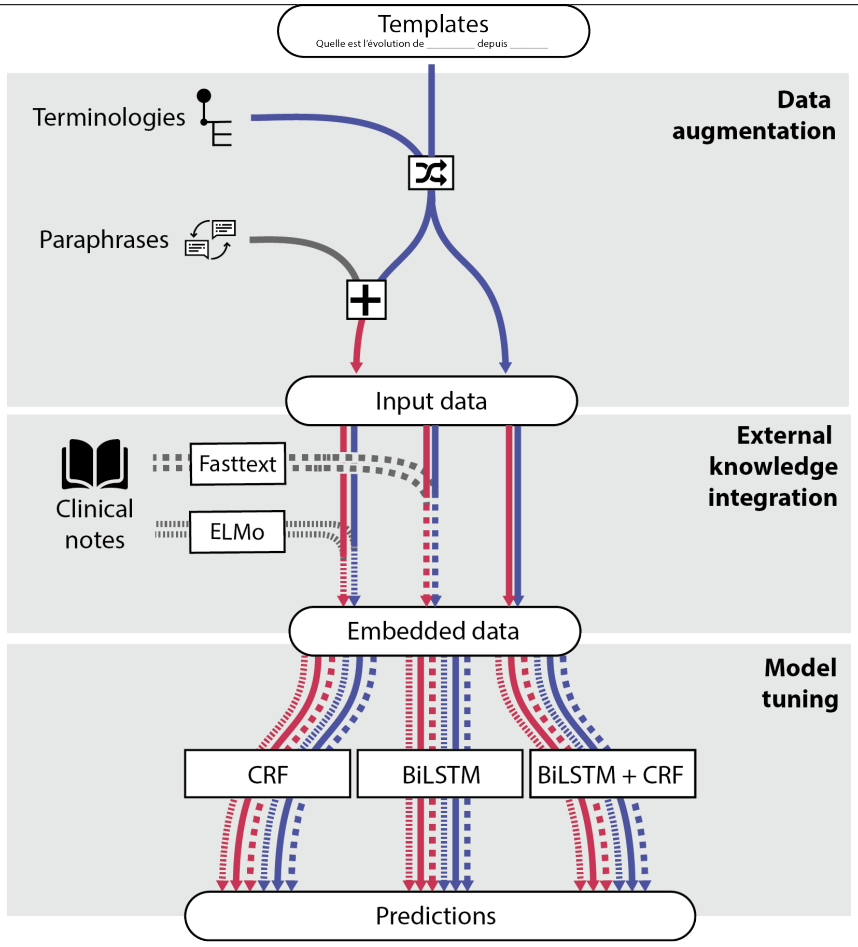


FIGURE 2.1 : Shcéma expérimentatl

2.3.1 Description de la tâche : module de compréhension de la langue pour une tâche de dialogue sur la requête dans les dossiers patients informatisés.

Le but est de permettre à un médecin de rechercher des informations sur les données d'un patient concernant les résultats d'examens biologiques. Les modalités de questions possibles sont diverses et variées, et permettre au clinicien de poser les questions en langue naturelle pourrait aider à accéder à l'information pertinente de manière plus rapide et plus efficiente. Par exemple, "*Quel était le dernier résultat de créatinine ?*" ou "*Comment a évolué la créatinine depuis la dernière hospitalisation ?*"

Devant l'absence de jeu de données d'entrainement disponible, nous avons d'abord élaboré un set de questions *a priori* qui nous paraissaient pertinentes pour un tel outil. Ce jeu de questions *a priori* nous a servi pour produire les modèles de questions pour la génération des jeux de données d'apprentissage (Section 2.3.2). Dans un deuxième temps, nous avons réalisé une enquête auprès des médecins de l'hôpital Necker-Enfants malades à Paris : nous leur avons demander de proposer des questions qu'ils pourraient poser à un système de ce type. Nous avons ainsi recueilli 178 questions auprès de 28 médecins différents que nous avons annotées manuellement pour créer un jeu de données de test *vie réelle*.(Annexe B.1)

Les tâches de compréhension de la langue sont habituellement divisées en 3 sous-tâches : classification du domaine, classification de l'intention de l'utilisateur (*intent*) et détection des entités (*slot filling*). Dans notre cas, le domaine est fixé, nous nous focalisons donc sur la détection de l'intention et la détection des entités. Une modélisation de la tâche est disponible dans la Figure 2.2.

La détection des entités consiste à repérer dans une utterance les tokens ou groupes de tokens correspondant à des entités d'intérêt. Par exemple, dans la question "*Comment a évolué la créatinine depuis la dernière hospitalisation ?*" "*créatinine*" est une entité de type résultat biologique et "*la dernière hospitalisation*" est une entité de type date. Voir Section 2.3.1

De manière générale, la classification de l'intention de l'utilisateur consiste en une unique tâche de classification multiclasse pour l'ensemble de l'utterance. Compte-tenu de la diversité des questions possibles, nous avons décidé d'adopter une stratégie de classification multi-axiale : la classification des intentions était ici découpée en 4 sous-tâches correspondant à 4 axes distincts permettant de reconstituer la diversité des possibles. Voir Section 2.3.1

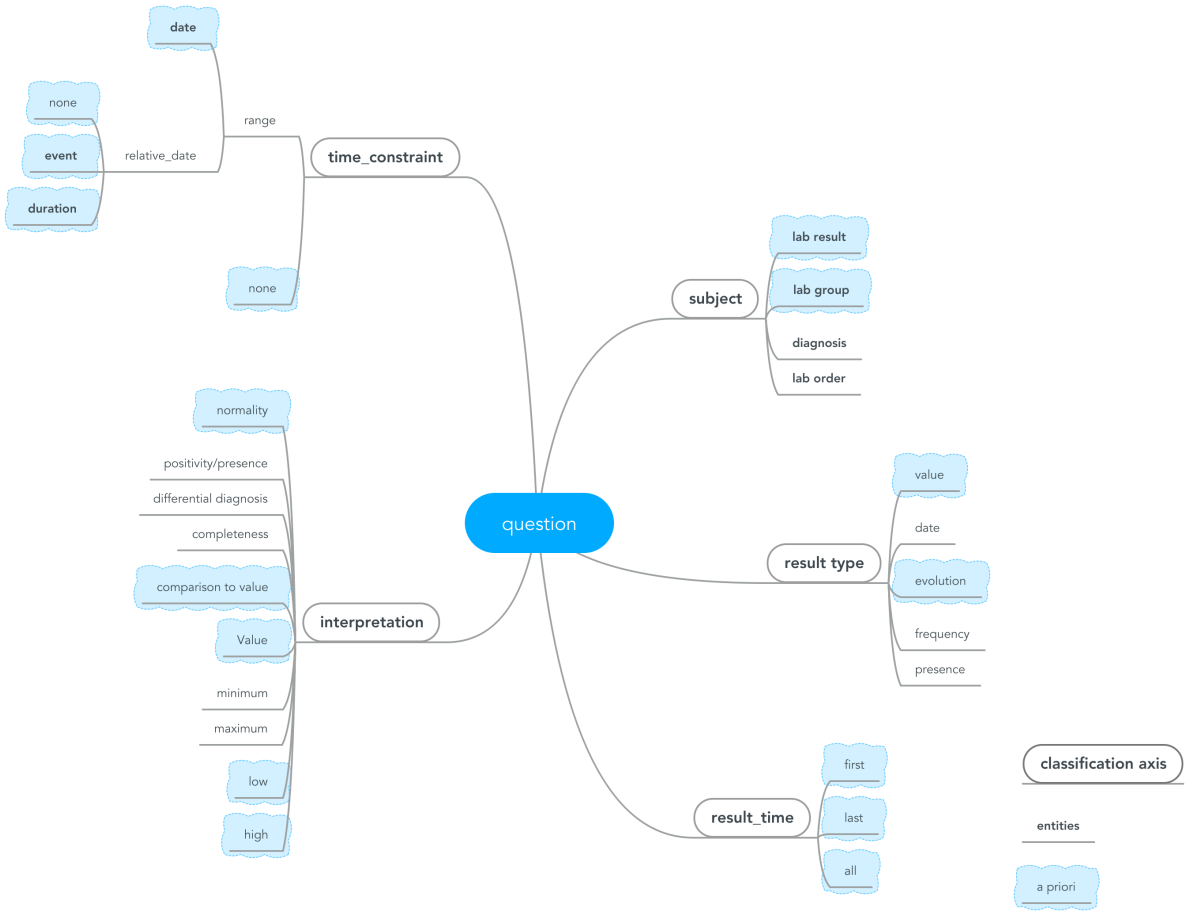


FIGURE 2.2 : Modélisation des questions relatives aux examens de biologie. en bleu, les modalités de questions envisagées a priori. en blanc, les autres modalités découvertes dans les questions réelles.

	Date	Mentions d'examen
mentions dans le jeu de test	34	177
Longueur médiane [min-max]	3 [0 - 6]	2 [1 - 11]
vocabulaire dans le jeu d'entraînement	1,364	451
vocabulaire dans le jeu d'évaluation (intersection)	58(0.38)	250(0.28)

TABLE 2.1 : Description des mentions

Labellisation de séquence

Nous distinguons ici 2 types de labels (2.2) : les mentions d'examens (*lab mention*) (*e.g.*, créatinine, protéine C réactive, CRP, ionogramme sanguin) et les dates (*e.g.* "27/03/2015," "3 jours," "dernière hospitalisation"). Les mentions d'examens peuvent correspondre à des examens uniques ou des groupes d'examens. De plus, les dates peuvent correspondre à des dates réelles, des durées ou des dates relatives à un jour ou un évènement. Il peut s'agir de dates fixes ou d'intervalles de dates. En cas d'intervalle entre 2 dates fixes, les 2 dates sont identifiées dans l'uterrance. Dans le jeu de données d'entraînement (voir section 2.3.2), il y avait 336 mentions d'examens distinctes, avec une longueur variant de 1 à 11 tokens et une longueur médiane de 2 tokens. Il y avait 28% de recouvrement entre le vocabulaire du jeu d'entraînement et le vocabulaire du jeu de test (vie réelle). Concernant les dates, la longueur était plus stable avec une médiane à 3, allant de 0 à 6 tokens. Le recouvrement de vocabulaire était de 38% entre le l'entraînement et le test. (Table 2.1)

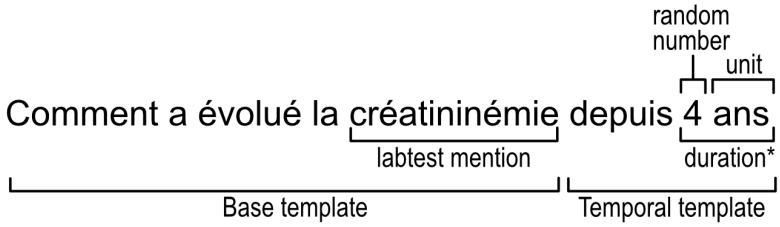
Classification.

Il y a 4 sous-tâches, chacune représentant un axe distinct (Figure 2.2) :
- le type (5 catégories, e.g. valeur, évolution, date, présence, aucun) ;
- l'interprétation (5 catégories, e.g. normalité, valeur, bas, haut, présence) ;
- la temporalité (3 catégories, e.g. premier, dernier, tous) ;
- les contraintes temporelles (4 catégories, e.g. intervalle, date, nombre, aucun).

La combinaison de ces axes permettant de reconstituer en post-traitement l'intention complète de l'utilisateur. Ainsi, *est-ce que la créatinine a augmenté ?* sera classé en [type:"évolution," interprétation:"haut," temporalité:"tous," contraintes:"aucune"] La modélisation de la temporalité est divisée en deux axes différents : d'une part un axe contrainte qui va restreindre la requête à un intervalle de temps donné. Cet intervalle peut être défini explicitement avec des dates fixes ou des dates relatives, ou être implicite et défini par une durée depuis ou jusqu'à une certaine date ou évènement. Une fois cette contrainte temporelle résolue, le deuxième axe de temps, l'axe "temporalité" agit comme un filtre sur les résultats à retenir au sein de cet intervalle. On pourra ainsi conserver le premier élément, le dernier élément ou l'ensemble des éléments pour composer une réponse. Les deux autres axes concernent le type et l'interprétation du résultat. Le type se rapporte au type d'information sur les examens biologiques que l'utilisateur souhaite avoir : il peut s'agir de la valeur du résultat, de la date réalisation, de la présence d'un élément dans le résultat, ou de l'évolution des résultats. L'interprétation va transformer l'élément sélection dans l'axe type : si le type était valeur et que l'interprétation est normalité, la réponse attendue concernera la normalité du résultat de l'examen.
Le raisonnement derrière cette approche en sous tâches est de diviser la complexité d'une classification multi-classes avec un nombre de classes élevé (180 classes ici) en une combinaison de plusieurs tâches de classification moins difficiles et plus généralisables.

2.3.2 Génération de données

Etant donnée l'absence de jeu de données disponible pour cette tâche, nous avons généré un jeu de données en utilisant un générateur dédié. Inspirés par le travail de Bordes *et al.* [145], nous avons développé des modèles de questions (templates) : 223 pour la base de la question (e.g. "quel est le résultat du dernier "), 23 modèles de modifieurs temporels (e.g. "depuis <date|durée|évènement>") et une liste de 409 mentions de résultats de tests de laboratoire (e.g. créatinine, hémoglobine, CRP, NFS). Chaque question générée associant aléatoirement un modèle de base, un modifieur temporel et une mention d'examen. (Figure 2.3)



Base template	170
Temporal templates	23
Labtest mentions	336
Duration units	15
Random duration range	[1-30]
Date range	[01/01/1950 - 31/12/2017]
Events	17

FIGURE 2.3 : Combinaison des modèles et modifieurs pour le jeu de données d'entrainement.

2.3.3 Augmentation de données par paraphrase

Les chercheurs travaillant sur des méthodes gourmandes en données, comme les réseaux de neurones, font régulièrement appel à la génération de paraphrase, y compris pour des tâches de question-réponse[146]. Nous encourageons le lecteur à se référer aux revues de la littérature disponibles concernant la génération de paraphrase [147,148], ainsi qu'aux récentes avancées utilisant les approches neuronales [149]. Une approche consiste à utiliser la Paraphrase Database (PPDB)[150], une grande base de données multilangues de paires de paraphrases (plus de 100 millions de paires en anglais) avec des variations lexicales, syntaxiques et de formulation. Une autre approche s'appuie sur la traduction automatique [151]. Par exemple, Zhang et al. [152] a utiliser une méthode de traduction par langage pivot pour dériver des paraphrases de mots clés dans des questions. La méthode par langage pivot consiste à traduire une première fois de la langue A vers la langue B puis traduire une nouvelle dois pour revenir à la langue A.

Dans ce travail, nous avons utilisé une méthode basée sur la traduction automatisée pour augmenter la variabilité du jeu de données d'entraînement en produisant des paraphrases des modèles de questions (définis dans la section précédente). Nous avons traduit chaque phrase vers un ou plusieurs langues (langues pivots) et traduit de nouveau vers la langue source.

Nous nous sommes servis pour cela de l'API Google Translate [153]. Pour chaque modèle de

question, nous avons sélectionné aléatoirement 10 langues parmi les plus de 60 disponibles dans l'API. Pour chaque langue, nous avons réalisé une traduction en pivot et conservé les paraphrases uniques obtenues en supprimant les doublons. Nous avons pu ensuite ajouter les paraphrases obtenues au modèles de questions utilisées pour la génération.

2.3.4 Injection de connaissances latentes

Les récentes avancées en représentation du langage comme les words embeddings statiques (e.g. word2vec [30], Glove [31], fastText [32]) et les embeddings contextuels (*e.g.* ELMo [33], BERT[34]) ont permis une amélioration significative des performances en traitement automatique de la langue en général et en compréhension de la langue en particulier. Ces méthodes tirent avantage de grands corpus de texte pour apprendre, en se basant sur la distribution des mots et leur contexte d'utilisation, une représentation des mots dans un espace vectoriel dense s'appuyant sur leur utilisation sémantique et syntaxique [30]. Un des avantages de ces approches est qu'elles sont auto-supervisées : il n'y a pas besoin d'annoter manuellement de jeu de données puisqu'elles ne nécessitent que du texte brut. En revanche, le corpus d'entraînement doit être suffisamment grand pour permettre l'apprentissage d'une représentation correcte du vocabulaire. Par exemple, il est d'usage d'utiliser l'ensemble des articles de Wikipedia® [154] pour entraîner les modèles [30–33]. Dans le domaine biomédical, le vocabulaire est très spécifique. De plus, l'observation attentive des textes cliniques dans les dossiers patients, nous montre une organisation, une syntaxe et un orthographe très différentes d'articles comme ceux que l'on peut trouver sur Wikipedia. En 2018, une étude de Wang *et al.* [155] a évalué les performances des embeddings word2vec et Glove, entraînés sur Wikipedia, sur la littérature biomédicale et sur des textes cliniques, sur différentes tâches de traitement automatique de la langue en anglais. Ils ont montré que l'entraînement sur les textes cliniques produisait de meilleures performances. Est-ce que ces résultats restent vrais avec des embeddings contextuels tels que ELMo ?

Dans cette étude, nous avons comparé 3 types d'embeddings :

- modèle skip-gram continu entraîné simple (baseline)
- modèle skip-gram continu avec information sur les sous-mots (i.e. chaque mot est représenté comme un sac de n-grams de caractères), tel qu'implémenté dans fastText [32]
- embeddings par modèles de langue (embeddings from language models, ELMo) dans lequel les vecteurs sont appris à partir des états latents d'un modèle de langue profond et bidirectionnel tel que décrit dans Peters *et al.* [33].

Nous avons également comparé les performances de ces méthodes en fonction du corpus sur lequel elles ont été entraînées : domaine général ou spécialisé. Pour le domaine général (ci-après, Wiki), c'était la version française de Wikipédia plus le jeu de données CommonCrawl en français. Pour ce jeu de données, nous avons utilisé les modèles pré-entraînés pour le français disponibles sur les sites internet de fastText et ELMo. Le jeu de données spécialisées (ci-après, EHR) était constitué d'un échantillon aléatoire d'un million de textes cliniques issus d'entrepôt de données clinique de l'hôpital Necker-Enfants malades à Paris [156]. Ce jeu de données comprenait 162M de tokens et un vocabulaire de taille 92k. Pour ce jeu de données, nous avons entraîné les modèles à l'hôpital. fastText a été entraîné avec 300 dimensions et une fenêtre de 5 tokens, et les autres paramètres par défaut. Pour ELMo, nous avons conservé uniquement un sous-ensemble de ce jeu de données (24M de tokens) compte-tenu du temps d'entraînement élevé et des ressources disponibles. Pour rester comparables entre les embeddings des différentes sources, nous avons concervé les hyper-paramètres décris dans Peters *et al.* [33]

	Enoncés	Modèles	mentions d'examen	Mots(*)	MHV†(*)	Perplexité(*)
Entraînement	16,000	170	336	144,850(140,492)	-	-
Développement	4,000	53	73	36,211(36,211)	4,724(4,544)	137.5(171.1)
Evaluation	178	-	-	1,579	467(390)	194.5(240)

* avec paraphrase

† Mots Hors Vocabulaire

TABLE 2.2 : Charactéristiques des jeux de données

2.4 Expériences

Nous avons séparé les modèles de questions et le mentions d'examens en deux sets distincts : entraînement (170 modèles, 336 mentions) et développement (53 modèles et 73 mentions) utilisé pour l'ajustement des hyper-paramètres. Le but de garder des modèles et des mentions spécifiquement pour le développement étant de favoriser les modèles de NLU avec de meilleures propriétés de généralisation. Le jeu de données de test, pour évaluer les performances finales étant le jeu de données "vie réelle" de 178 questions.

Nous avons généré 16 000 questions pour l'entraînement et 4000 pour le développement. Ensuite, nous avons effectué l'augmentation des modèles de question avec les paraphrases. Puis nous avons généré un deuxième lot de 16000 questions pour l'entrainement et 4000 pour le développement en incluant les modèles de question paraphrasés. (Table 2.2). Une façon usuelle de produire des systèmes de NLU spécialisés est d'élaborer des algorithmes basés sur les règles pour réaliser le découpage sémantique des énoncés des utilisateurs [157]. Cependant, développer de tels systèmes est très consommateur de temps et souvent difficile à maintenir. Les systèmes moderne de NLU utilisent préférentiellement des systèmes d'apprentissage statistique pour réaliser cette tâche [158]. Avant l'avènement des systèmes à base de réseaux de neurones, l'état de l'art était occupé par des systèmes de types conditional random fields (CRF) [10]. Aujourd'hui, ces systèmes ont été surpassés par les approches basées sur des réseaux de neurones comme les réseaux de neurones convolutionels (CNN) ou les réseaux de neurones récurrents (RNN). Sur la tâche de labellisation de séquence, les RNNs et plus spécifiquement les LSTM (long short term memory units) [18] sont les plus utilisés. L'état de l'art étant occupé par une combinaison des LSTM bidirectionnels et des CRF, les BiLSTM-CRF [124].

Par évaluer la capacité de ces modèles à généraliser à de nouvelles données, nous avons évalué trois types de modèles pour la tâche de labellisation de séquence : CRF, BiLSTM et BiLSTM-CRF. La couche d'entrée était nourrie par les questions générées à partir des modèles de question standards ou augmentés des modèles paraphrasés. Nous avons également comparés avec les différents embeddings décrits précédemment.

Pour tous les modèles (à l'exception de ceux basés sur ELMo) nous avons ajouté des features standards en entrée : lemmes normalisés et part-of-speech (POS) tagging. La partie labellisation de séquence était constituée d'un CRF seul ou de deux couches de biLSTM ou de deux couches de BiLSTM suivies d'un CRF. Les paramètres de tuning étaient : la dimension des embeddings (50, 100, 300) sauf pour ELMo (fixé à 1024), le nombre d'unités dans le BiLSTM (64,128,256), les fractions de dropout après l'embedding et après le LSTM [0.1-0.5]. Concernant la partie classification, il s'agissait de modèles constitués par les mêmes embeddings, une couche de 1D-convolution avec des kernels de taille 2 à 5, 50 à 250 filtres et une activation ReLu. Cette couche était suivie par une couche de max-pooling et un fonction softmax. La fonction d'optimisation utilisée était Adam [159]. Tous les modèles ont été implémentés avec Keras[160] et un backend

Model	F1-score [95%CI][a]
CRF + para	.37 [.35-.39]
CRF	.43 [.42-.44]
BiLSTM	.59 [.58-.61]
BiLSTM + para	.66 [.65-.68]
BiLSTM + CRF	.62 [.60-.63]
BiLSTM + CRF + para	.60 [.59-.62]
CRF + FastText + para	.62 [.60-.63]
CRF + FastText	.62 [.61-.64]
BiLSTM + CRF + FastText	.67 [.65-.68]
BiLSTM + CRF + FastText + para	.67 [.65-.69]
BiLSTM + FastText	.69 [.67-.71]
BiLSTM + FastText + para	.68 [.67-.70]
BiLSTM + CRF + ELMO + para	.73 [.71-.74]
BiLSTM + ELMO + para	.74 [.72-.75]
CRF + ELMO	.75 [.73-.76]
CRF + ELMO + para	.75 [.74-.76]
BiLSTM + CRF + ELMO	.76 [.74-.77]
BiLSTM + ELMO	.77 [.76-.79]

[a] para = paraphrase

TABLE 2.3 : Résultat des expériences sur la tâche de labellisation de séquence

Tensorflow [161].

Tous les résultats sont rapportés en terme de F-mesure pondérée, calculée sur 10 répétitions de 5x cross-validation sur le test set.

2.5 Résultats

Globalement, les meilleurs résultats sur la labellisation de séquence ainsi que sur la classification sont obtenus avec les modèles incluant les représentations ELMo pour injecter des connaissances externes.

Sur la tâche de labellisation de séquence, les modèles avec ELMo-BiLSTM et ELMo-BiLSTM-CRF entraîné sur les données EHR ont obtenu un F1-score de 0.76(IC95% [0.74-0.77]) et 0.77 (IC95% [0.76-0.79]) respectivement. (voir Table 2.3, Figure 2.4). Sur la tâche de classification, les meilleurs résultats sont obtenus avec ELMo sur trois des quatres sous-tâches et à égalité avec fastText-paraphrases sur la quatrième. (voir Table 2.4, Figure 2.5)

	Sub-task (F1 score)			
Modèle	Type	Interprétation	Temporalité	Contrainte Temporelle
Normal				
Embedding simple	.64 [.62-.67]	.65 [.63-.68]	.68 [.65-.70]	.40 [.38-.42]
EHR ELMO	.70 [.68-.73]	.64 [.62-.67]	.77 [.75-.79]	.41 [.40-.44]
EHR FastText	.69 [.67-.72]	.68 [.66-.70]	.70 [.68-.72]	.42 [.40-.44]
Paraphrases				
Embedding simple + para	.62 [.59-.64]	.71 [.69-.72]	.72 [.70-.74]	.53 [.50-.55]
EHR FastText + para	.64 [.62-.66]	.68 [.66-.70]	.74 [.72-.76]	.72 [.69-.74]
EHR ELMO + para	.65 [.62-.67]	.68 [.65-.70]	.75 [.73-.77]	.72 [.70-.74]

TABLE 2.4 : Résultats des expériences sur la tâche de classification

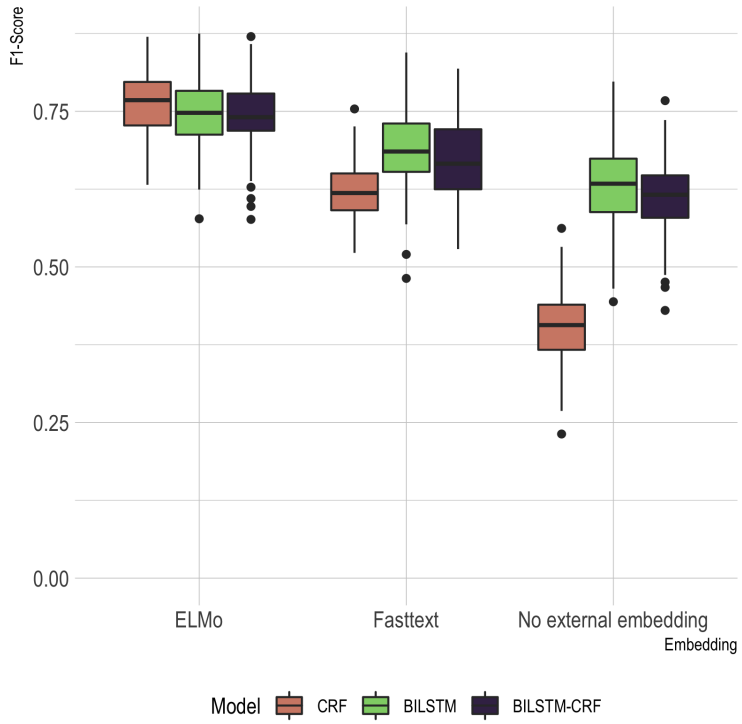


FIGURE 2.4 : Labellisation de séquence

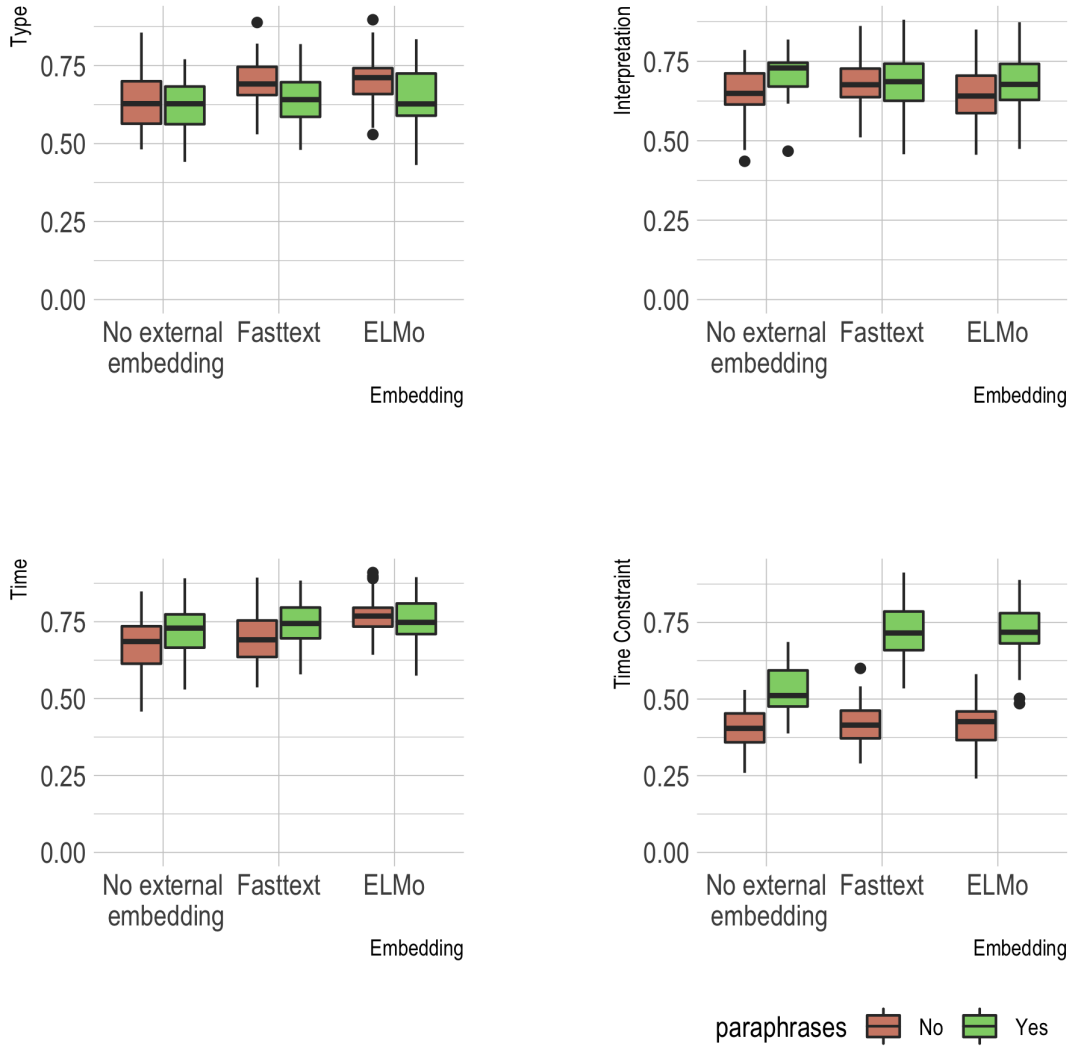


FIGURE 2.5 : Classification

Sur la labellisation de séquence, ajouter des connaissances latentes avec fastText ou ELMo entrainés sur EHR augmente la généralisabilité des modèles indépendamment de l'architecture du modèle en aval. Entraînés sur EHR, les modèles avec ELMo avaient un F1-score moyen de 0.75(±0.05) et avec fasText de 0.66(±0.06) contre 0.55(±0.12) sans embedding externe. Les embeddings entraînés sur Wiki amélioraient la baseline mais restaient moins performants que ceux entraînés sur EHR.(Table 2.5). Il est intéressant de noter que les résultats de ELMo et de fastText entraînés sur Wiki sont équivalents à 0.67 IC95%[0.67-0.70].

L'ajout de paraphrases aux modèles de questions n'a pas amélioré les résultats sur cette tâche et a même tendance à abaisser les performances : ELMo 0.76(±0.05) sans paraphrases et 0.74(±0.05) avec ; fastText 0.66(±0.06) versus 0.66(±0.06) ; sans embedding externe 0.55(±0.10) versus 0.54(±0.14). Concernant le type de modèle, les Bi-LSTM sans ou avec CRF, on obtenu de meilleures performances que les CRF seuls avec des F1-scores de 0.69(±0.08), 0.68(±0.08) et 0.59(±0.15), respectivement. Sur la tâche de classification, nous avons également observé de meilleurs résultats avec ELMo et fastText entraînés sur EHR : F1-scores moyens 0.68(±0.12) avec ELMo, 0.66(±0.11) avec fastText et 0.61(±0.12) sans embedding externe. Sur cette tâche,

Method	Sequence labelling	Intent classification
	F1-score [95%CI]	Mean F1-score [95%CI]
Baseline (only training set)	0.62 [0.61-0.64]	0.62 [0.58-0.65]
Fasttext on Wiki	0.69 [0.67-0.70]	0.52 [0.49-0.55]
Fasttext on EHR	0.67 [0.61-0.73]	0.66 [0.63-0.69]
ELMo on Wiki	0.69 [0.67-0.70]	0.49 [0.46-0.52]
ELMo on EHR	0.76 [0.74-0.77]	0.70 [0.67-0.73]

TABLE 2.5 : Comparaison des différents types d'embedding

les modèles entrainés sur Wiki ont obtenu des résultats faibles avec un F1-score à 0.52 (IC95% [0.49-0.55]) pour fasText et 0.49 (IC95% [0.46-0.52]) pour ELMo inférieurs à la baseline.(Table 2.5) Contrairement à la tâche de labellisation, ajouter des paraphrases tend à donner de meilleurs résultats avec des F1-score à 0.63($\pm$0.13) sans paraphrase et 0.67($\pm$0.10) avec. (Figure 2.5).

2.6 Discussion

Il est intéressant de noter que les résultats obtenus avec les meilleurs modèles sur chaque tâche montrent qu'il est possible d'utiliser cette méthode pour proposer un système baseline pour des tâches de compréhension du language en l'absence de données pré-existentes.

Nos résultats, non seulement confirment ceux observés par Wang *et al* [155] sur l'intérêt d'incorporer des connaissances externes en utilisant un grand corpus spécifique du domaine, mais mettent également en évidence l'intérêt d'utiliser des modèles de langue à la place d'embeddings statiques pour incorporer ces connaissances. Dans notre étude, les résultats utilisant ELMo étaient systématiquement meilleurs qu'avec fastText bien que les modèles aient été entraînés sur les mêmes données. Cela provient probablement d'une meilleure représentation du contexte avec ELMo qu'avec fastText. fastText prend en compte les tokens dans une certaine fenêtre et ne considère pas leur séquence. ELMo est un modèle de langue qui considère l'ensemble du contexte de l'utilisation du token pour générer une représentation spécifique du token dans ce contexte. Il est à noter que la tâche de labellisation utilisée ici n'est pas très complexe compte-tenu du nombre de catégories d'entités à identifier. Refaire cette analyse avec une tâche plus complexe pourrait donner des résultats différents.

La modélisation des questions que nous avons proposé ici a des limites. Parmi les questions que nous avons recueilli auprès des médecins de Necker, certaines se sont avérées difficiles, voir impossible à modéliser avec une valeur unique par axe. En effets nous avons constaté des questions complexes qui demandaient parfois plusieurs entités à la fois (*Suivi des leucocytes/lymphocytes/monocytes sur 1 semaine*) plusieurs types de résultats pour une même entité (*quelle est la date et le résultat de la dernière biopsie rénale ?*, *Les RAI ont elles été faites et sont elles à jour ?*), une interprétation en deux étapes avec une résolution conditionnée par une première requête (*quel est le résultat bactériologique de la biopsie du 4/02/17 ?*). Dans le dernier exemple, il faut dans un premier temps résoudre la *biopsie du 4/02/17* avant de pouvoir rechercher le résultat de l'*examen bactériologique* qui s'y réfère. (Annexe B.1) Dans le cas de ces questions complexes, l'approche de modélisation à plat, séparant entités et intentions semble limitée. Récemment, des méthodes de modélisation hierarchique des intentions et des entités ont été proposées [162,163]. Ces méthodes s'appuient sur des travaux liés de Dyer *et al* [164] sur

la modélisation hiérarchique de la structure des phrases (recurrent neural network grammars, RNNG). L'idée est de modéliser, dans un même arbre hiérarchique, à la fois les intentions et les entités. Cela permet une représentation plus riche, avec notamment la possibilité de résoudre en plusieurs temps les questions complexes. Dans la Figure 2.6, un exemple d'utilisation de ce type de représentation. Les intentions sont représentées par les éléments `IN:` et les entités (slots) par les éléments `SL:`. Dans l'annexe C, nous proposons une représentation sous cette forme pour le jeu de données questions réelles selon le formalisme proposé par Gupta et al [162].

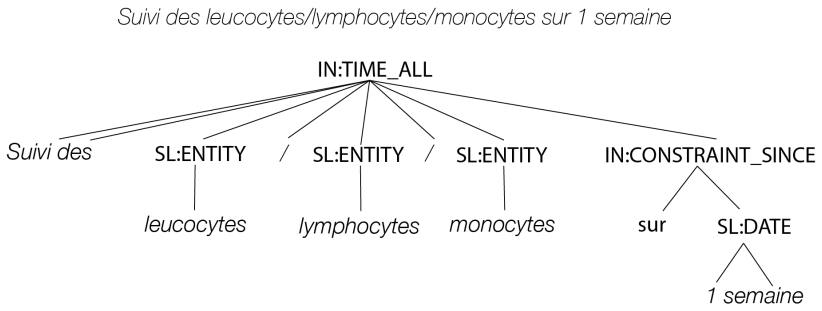


FIGURE 2.6 : Exemple de parsing utilisant une approche hiérarchique combinant intention et entités

Les meilleures améliorations provoquées par l'utilisation d'ELMo se sont manifestées sur la labellisation de séquence, suggérant que les tâches nécessitant des lexiques étendus bénéficieraient le plus des pre-embeddings. En effet, l'identification des mentions d'examen dépend fortement du vocabulaire disponible pendant l'entraînement, or nous avons vu que 72% des mentions du test set étaient absentes du jeu d'entraînement. Dans cette configuration, les représentations apprises sur un grand corpus spécialisé ont aidé le modèle à obtenir une meilleure généralisabilité.

Le test set montre 29% de mots hors vocabulaire au total. Cela peut expliquer en partie les améliorations observées sur la tâche de classification. Il est intéressant de noter que les deux embeddings entraînés sur Wiki ont obtenu de très mauvais résultats. Ce bruit pourrait venir de l'absence de vocabulaire spécialisé dans le corpus Wikipédia ou d'une utilisation contextuelle différente de ces mots par rapport au corpus de spécialité.

Sheikhshabbafghi et al. [165] a évalué les embeddings ELMo sur des tâches de reconnaissance d'entité nommées dans le domaine biomédical. Ils ont également observé des résultats améliorés avec l'utilisation de corpus dans le domaine (articles Pubmed) pour entraîner ELMo. Il serait intéressant d'évaluer des embeddings ELMo entraînés sur un corpus d'articles biomédicaux en français, mais collecter un tel corpus pourra s'avérer délicat.

Les embeddings contextuels ELMo ne sont plus à l'état de l'art aujourd'hui, les méthodes à base de transformers [27] sont désormais plus efficaces. Reproduire cette étude en utilisant des représentations de type BERT [34] ou CamemBERT [39] pourrait permettre d'améliorer les résultats.

D'autre part, les résultats sur les paraphrases sont plus difficiles à interpréter : ils sont légèrement meilleurs sur la tâche de classification mais pas sur la tâche de labellisation de séquence. Cela pourrait s'expliquer par la méthode de génération de paraphrase par traduction pivot. En effet, la qualité des paraphrases obtenues n'est peut-être pas suffisante pour la tâche. Utiliser des méthodes de génération de paraphrases plus sophistiquées pourrait permettre d'améliorer

ces résultats. Nous pouvons penser par exemple à l'utilisation d'architectures de traduction neuronales de types encodeur-décodeur [166] ou des modèle de génération de texte "retrieval based" [167]. Il est également envisageable d'utiliser des modèles de langues spécialisés dans la génération de texte tels que GPT-3 [168] et de les réentrainer à répondre à la tâche spécifique de paraphrase qui nous intéresse ici.

2.7 Conclusion

Les modèles de NLU entraînés à partir de données générées en utilisant la méthode proposée ont montré des performances intéressantes. Ces méthodes pourraient être utilisées pour apprendre un système baseline afin de bootstraper un système de dialogue et commencer à recueillir des donner d'utilisateurs finaux. Ainsi, en utilisant les modèles développés ici, il nous sera possible de mettre en place un premier agent conversationnel permettant aux cliniciens de requêter des informations sur les résultats d'examens biologiques. Il ne s'agit ici que de la première pierre vers un système de ce type, et tous les autres modules d'un système de dialogue opérationnel restent à développer. Plusieurs challenges ont été mis en évidences pendant cet étude. Tout d'abord, la diversité et la complexité des questions posées par des utilisateurs réels avait été sous estimées. Bien entendu, dans un domaine fermé comme ici, il est toujours possible de concevoir de nouveaux modèles de questions pour répondre aux modalités de questions qui n'avaient pas été anticipées. Cela étant dit, d'autres approches peuvent également être envisagées. Notamment la génération de questions basées sur les éléments présents dans les bases de données.[169] En effet, même si nous n'avions pas imaginé de questions sur les dates d'examen par exemple, le fait que cette information soit présente dans les bases de données pourrait permettre de générer des questions y afférant.

La complexité des questions est un sujet différent, nous avons déjà discuté la possibilité d'utiliser des systèmes hierarchiques pour modéliser les interconnexions entre intentions et entités. En revanche, un certain nombre de questions dans le jeu de données vie réelle se rapportait à des interprétations médicales de résultats plutôt qu'à leur valeur brute. Ainsi, nous demandait-on si le patient avait une anémie plutôt que son taux d'hémoglobine. Cela implique d'instiller une couche d'intelligence supplémentaire dans notre système, supposant que de telles connaissances existent dans une base formalisée. Or, au meilleur de nos connaissances, une telle base n'existe pas à l'échelle de la médecine en général. Il s'agira dons d'être capable dans un premier temps de construire une base de connaissance formelle reliant un diagnostic au résultats d'examens et informations pertinentes, en prenant soin d'y inclure les règles de décision quantifiées permettant de les relier.

Au niveau des représentations de mots, nous avons pu constater que les embeddings contextuels entrainés sur un jeu de données spécialisées permettait d'améliorer les performances des modèles de manière non négligeable. Nous aurons l'occasion de revoir dès le Chapitre 3 que cette observation s'applique également à d'autres types de tâches.

Ce travail a donné lieu à deux publications :

Une première dans le workshop Machine Learning for health de la conférence Neurips en 2018 à Montréal :

Neuraz, Antoine, Leonardo Campillos Llanos, Anita Burgun, and Sophie Rosset. 2018.
"Natural Language Understanding for Task Oriented Dialog in the Biomedical Domain in a Low Resources Context."
ArXiv:1811.09417 [Cs], November. http://arxiv.org/abs/1811.09417.

La deuxième publication, dans la revue Studies in Health Technology and Informatics, à l'occasion de la conférence MIE 2020 :

Neuraz, Antoine, Bastien Rance, Nicolas Garcelon, Leonardo Campillos Llanos, Anita Burgun, and Sophie Rosset. 2020.
"The Impact of Specialized Corpora for Word Embeddings in Natural Langage Understanding." Studies in Health Technology and Informatics 270 (June) : 432–36. https://doi.org/10.3233/SHTI200197.

2.8 *Article : Natural language understanding for task oriented dialog in the biomedical domain in a low resources context*

Natural language understanding for task oriented dialog in the biomedical domain in a low resources context

Antoine Neuraz, Anita Burgun
Department of Biomedical Informatics, Hôpital Necker-Enfants Malades, APHP
INSERM UMRS 1138, Team 22, Paris Descartes, Université Sorbonne Paris Cité
{antoine.neuraz,anita.burgun}@aphp.fr

Leonardo Campillos Llanos, Sophie Rosset
LIMSI, CNRS, Université Paris Saclay, France
{leonardo.campillos,sophie.rosset}@limsi.fr

Abstract

In the biomedical domain, the lack of sharable datasets often limit the possibility of developing natural language processing systems, especially dialogue applications and natural language understanding models. To overcome this issue, we explore data generation using templates and terminologies and data augmentation approaches. Namely, we report our experiments using paraphrasing and word representations learned on a large EHR corpus with Fasttext and ELMo, to learn a NLU model without any available dataset. We evaluate on a NLU task of natural language queries in EHRs divided in slot-filling and intent classification sub-tasks. On the slot-filling task, we obtain a F-score of 0.76 with the ELMo representation; and on the classification task, a mean F-score of 0.71. Our results show that this method could be used to develop a baseline system.

1 Introduction

There is a growing research interest on conversational interfaces for biomedical natural language processing [Laranjo et al., 2018]. Dialogue systems involve several components [Jokinen and McTear, 2009]: a natural language understanding (NLU) module, a dialogue manager, a generation module and a module for querying the database. We are interested here in the NLU component, which allows the system to understand user's utterances through the semantic analysis and formalization of queries.

To develop a NLU model using a machine learning approach, one of the first requirements would be a training dataset. This dataset requires user utterances (input of the NLU), along with a formal representation (output of the NLU). The dataset needs to be large enough to be representative of the task. But what if this dataset does not exist? Depending on the task and the language, it is likely that one can not find any suitable training dataset. The biomedical domain is a good use case when it comes to low resources in terms of data, especially in languages other than English. Due to privacy issues, it is difficult to share real world medical data. Dialog systems in the medical domain have been applied for patient counselling on a wide range of topics, from medical conditions to medication intake [Azevedo et al., 2018]. In most of the systems, interactive capabilities are based on limited, constrained natural language input: for example, users are presented with a menu of multiple-choice questions. In contrast, dialogue systems allow users to access data in a much more natural way—through speech or typed input.[Laranjo et al., 2018].

Machine Learning for Health (ML4H) Workshop at NeurIPS 2018.

One solution to overcome the absence of training data is to generate a training dataset based on a few examples and augment it using known terminologies, external knowledge and paraphrases. In this paper, we assess how well models created using such a generated training set perform on real world data and compare their performances on biomedical NLU task in French.

2 Data generation and augmentation

In this section, we describe the task and the methods we use to generate the training and development sets. We explain the methods for generating data using templates and terminologies, generating paraphrases of templates using pivot translations and incorporating external knowledge with word embeddings and language models. Figure S2 details the general schema of this work.

2.1 Description of the task: NLU in a dialogue task to query EHRs

The aim of the task is to perform Natural Language Understanding of user input. This step will enable physicians to perform queries in Electronic Health records (EHRs) in natural language. The set of queries that a physician may have about characteristics and results of a patient is broad and diverse, therefore, enabling queries in natural language may help accessing information more efficiently. For this purpose, we asked to medical doctors from a French university hospital some examples of questions they would ask to a dialog system aimed at querying information about biological tests results. We collected a set of 178 questions that we annotated manually as a gold standard.

NLU tasks can usually be divided into 3 sub-tasks: domain classification, intent classification, and slot-filling [Tur and Mori, 2011]. For this task in a restricted domain (bio-medicine), we focus on the two latter: slot-filling (sequence labelling) and intent classification.

Sequence labelling. We distinguish two types of labels: lab mentions (e.g. "créatinine" *creatinin*, "protéine C réactive" *C reactive protein*) and dates (e.g. "27/03/2015", "depuis 3 jours" *since 3 days*). In the training set (generated data, see 2.2), the number of distinct lab mentions is 336, for a length ranging from 1 to 11 tokens and a median length of 2. There is 28% of overlapping between the vocabularies of the train set and the test set (real world data). Regarding the date labels, they include actual dates, relative dates and time ranges. The length is more stable with a median of 3, ranging from 0 to 6 tokens. The vocabulary overlap is 38% between the train and test sets. (see Table S1)

Classification. There are 4 sub-tasks representing 4 axes of classification. For each utterance, we assign one label per axis. Two axes concern the results of the lab exams (i.e. the type of result (5 categories, e.g. value, evolution, date) and interpretation of the result (5 categories, e.g. normality, value, low, high, presence). The two latter concern temporal aspects (i.e. the time of result (3 categories, e.g. first, last, all) and constraints on time (4 categories, e.g. none, range, date, number).

2.2 Data generation

Given the lack of suitable dataset for training, we generate a training dataset using a tailored generator. Inspired by Bordes et al. [2016], we developed questions templates: 223 for the core of the question (e.g. "quel est le résultat du dernier <lab mention>", *what is the result of the last <lab mention>*), 23 temporal modifier templates (e.g. "depuis <date|duration|event>" *since <date|duration|event>*) , and a list of 409 mentions of laboratory tests results (hereafter, lab mentions, e.g. "créatinine" *creatinin*, "hémoglobine" *hemoglobin*). Each generated question randomly associates a base template, a temporal modifier template and a lab mention to create unique questions (see Figure S1).

2.3 Data augmentation with paraphrases

Researchers working with data-intensive methods (such as neural networks) already resort to the generation of paraphrases, even for question-answering tasks [Dong et al., 2017]. We refer the reader to available reviews on methods of paraphrase generation [Androutsopoulos and Malakasiotis, 2010, Madnani and Dorr, 2010], including recent advances using neural approaches [Iyyer et al., 2018], and focusing on methods applied for paraphrasing questions. A recent approach makes use of the Paraphrase Database (PPDB) [Ganitkevitch et al., 2013], a large multilingual collection of paraphrase pairs (over 100 million pairs for English) with lexical, syntactic and phrasal variations. Another method relies on machine translation [Duboue and Chu-Carroll, 2006]. For example, [Zhang et al.,

2015] derive paraphrases of key words in questions by translating them to a pivot language (they experimented with 11 languages), then back to the source language.

We use a machine translation method to increase the variability of the training set by producing paraphrases of the question templates. We translate each sentence to one or several languages (pivot languages) and translate the result back to the source language. We used the Google Translate API [Google] for this step. For each template, we randomly select 10 of the 60+ languages available in the API. For each language, we perform the pivot translation and kept the unique paraphrases obtained. We add the paraphrases to the set of templates used for the generation of the datasets.

2.4 Incorporating latent knowledge

Using embeddings of words learned on a large domain specific dataset of unlabeled data can be an effective source of latent knowledge [Wang et al., 2018]. We use one million of clinical notes from the clinical data warehouse of a local hospital in France. Leveraging this corpus, we compare three types of method: 1) word embeddings (continuous skip-gram) only on the training set (without external knowledge) as a baseline; 2) continuous skip-gram model with sub-word information (i.e. each word is represented as a bag of character n-grams), as implemented in Fasttext [Bojanowski et al., 2016], 3) embeddings from language models (ELMo) where the vectors are learned from the internal states of a deep bidirectional language model as described in Peters et al. [2018].

3 Experiments

We split the question templates and the lab mentions into two sets: training (170 templates and 336 mentions) and development (53 templates and 73 mentions). From each, we create two datasets by generating paraphrases (see section 2.3). We generate 16,000 utterances for the training set (80%) and 4,000 for the development set (20%) using templates without paraphrases, and the same quantities for the sets with paraphrases (Table S2). The test set (real world data) is kept aside for the evaluations.

A usual way of producing specialized NLU systems is to elaborate rule-based algorithms to perform the semantic parsing of user's utterances [Weston et al., 2015]. However, developing such system can be time consuming and is often difficult to maintain. Most of modern NLU systems use statistical learning models to perform this task [Young et al., 2013]. Before the raising of neural based systems, state of the art systems used conditional random fields (CRF) [Lafferty et al., 2001]. Nowadays, these systems tend to be outperformed by neural based approaches, such as convolutional neural networks (CNN) and recurrent neural networks (RNN). On the task of sequence labelling, RNNs and more specifically long short term memory units (LSTM) [Hochreiter and Schmidhuber, 1997] are the most used. More recent work combine bidirectional LSTMs (biLSTM) and CRF [Lample et al., 2016].

To assess the capacity of the models to generalize to new data, we evaluate three types of models for this task: CRF, bidirectional LSTM (biLSTM), and a combination of biLSTM and CRF [Lample et al., 2016]. The input layer is fed with the generated questions from templates only or from templates with paraphrases. The embeddings are learned either directly on the training set (no external knowledge), or on clinical notes using Fasttext or ELMo. For each combination, we test three different models: CRF, biLSTM and biLSTM+CRF. The details of the models and the tuning parameters are described in the supplementary materials (section S1). All the results are reported in terms of weighted F-measure, computed using 10 repetitions of five fold cross-validation over the test set.

4 Results and discussion

Overall, the best results on sequence labelling and classification tasks are obtained with the models including ELMo representations as the embeddings used to inject external knowledge. On the sequence labelling task, models with ELMo-biLSTM and ELMo-biLSTM-CRF obtained a F1-score of 0.76(95%CI [0.74-0.77]) and 0.77 (95%CI [0.76-0.79]) respectively (see Table 1, Figure S3). On the classification task, the best results are obtained with ELMo on three of the four sub-tasks and with Fasttext-paraphrases on the forth one (see Table 2, Figure S4).

On the sequence labelling task, adding latent knowledge with FastText or ELMo using a million clinical records increases the generalizability of the models regardless of the type of the downstream model. Models with ELMo have an average F1-score of $0.75(\pm 0.05)$, with FastText $0.66(\pm 0.06)$ and

without external knowledge $0.55(\pm 0.12)$. Adding paraphrases to the templates does not improve the results on this task and even tends to lower the results: ELMo $0.76(\pm 0.05)$ without paraphrases and $0.74(\pm 0.05)$ with; FastText $0.66(\pm 0.06)$ versus $0.66(\pm 0.06)$; no external embedding $0.55(\pm 0.10)$ versus $0.54(\pm 0.14)$. Regarding the type of model, biLSTM and biLSTM-CRF perform better than CRF only with F1-scores of $0.69(\pm 0.08)$, $0.68(\pm 0.08)$ and $0.59(\pm 0.15)$, respectively.

On the classification tasks, we also observe better results with ELMo and fastText than without external embedding: mean F1-scores of $0.68(\pm 0.12)$ with ELmo, $0.66(\pm 0.11)$ with FastText and $0.61(\pm 0.12)$ without external embedding. Unlike for the sequence labelling task, adding paraphrases to the training set tends to give better results with F1-scores of $0.63(\pm 0.13)$ without and $0.67(\pm 0.10)$.

Interestingly, the results obtained with the best models on each task show that it is possible to use our method to provide a baseline system for NLU tasks in the absence of a pre-existing data. Our results not only confirm those by Wang et al. [2018] regarding the interest of incorporating external knowledge using a large domain specific corpus; our outcomes also highlight the interest of using language models instead of only embeddings to incorporate this knowledge. In our study, the results using ELMo are systematically better than those with FastText although the models were learned on the same data. This may come from the better representation of the context in ELMo compared to FastText. FastText takes into account the tokens in the specified window, which can be described as a "bag of context". But ELMo is a language model and considers the full context of a token (at the sentence level). Of note, this sequence labelling task is not very complex, given the number of different labels. The results on a task with more labels might be lower. Moreover, the results with the paraphrases are more difficult to interpret: they are slightly better on the classification tasks but not on the sequence labelling task. This might come from the method of pivot translation used for producing this paraphrases. Indeed, the quality of the produced paraphrases may not be sufficient for the task. Using more sophisticated methods of paraphrasing could lead to different results.

5 Conclusion

NLU models learned on the data generated using the proposed method achieve interesting performances. These methods can be considered to learn a baseline model allowing to bootstrap a dialog system and start collecting data from end users. We are interested in exploring to which extent other sources to train embeddings (e.g. medical, non-clinical texts) yield similar results. It would also be interesting to conduct similar experiments in related tasks where data are scarce (e.g. NLU in dialogue systems for patient counselling or virtual patients).

Model	**F1-score [95% CI]**
CRF + para	.37 [.35-.39]
CRF	.43 [.42-.44]
BiLSTM	.59 [.58-.61]
BiLSTM + para	.66 [.65-.68]
BiLSTM + CRF	.62 [.60-.63]
BiLSTM + CRF + para	.60 [.59-.62]
CRF + FastText + para	.62 [.60-.63]
CRF + FastText	.62 [.61-.64]
BiLSTM + CRF + FastText	.67 [.65-.68]
BiLSTM + CRF + FastText + para	.67 [.65-.69]
BiLSTM + FastText	.69 [.67-.71]
BiLSTM + FastText + para	.68 [.67-.70]
BiLSTM + CRF + ELMO + para	.73 [.71-.74]
BiLSTM + ELMO + para	.74 [.72-.75]
CRF + ELMO	.75 [.73-.76]
CRF + ELMO + para	.75 [.74-.76]
BiLSTM + CRF + ELMO	**.76 [.74-.77]**
BiLSTM + ELMO	**.77 [.76-.79]**

Table 1: Results of the experiments for the sequence labelling task. para = paraphrases

Model	**Sub-task (F1-score)**	
	Type	Interpretation
train-set embedding	.64 [.62-.67]	.65 [.63-.68]
EHR ELMO	**.70 [.68-.73]**	.64 [.62-.67]
EHR FastText	69 [.67-.72]	.68 [.66-.70]
train-set + para	.62 [.59-.64]	**.71 [.69-.72]**
EHR FastText + para	.64 [.62-.66]	.68 [.66-.70]
EHR ELMO + para	.65 [.62-.67]	.68 [.65-.70]

Model	**Sub-task (F1-score)**	
	Time	Time constraint
train-set embedding	.68 [.65-.70]	.40 [.38-.42]
EHR ELMO	**.77 [.75-.79]**	.41 [.40-.44]
EHR FastText	.70 [.68-.72]	.42 [.40-.44]
train-set + para	.72 [.70-.74]	.53 [.50-.55]
EHR FastText + para	.74 [.72-.76]	**.72 [.69-.74]**
EHR ELMO + para	.75 [.73-.77]	.72 [.70-.74]

Table 2: Results of the experiments for the classification task. para = paraphrases

References

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

I. Androutsopoulos and P. Malakasiotis. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187, 2010.

R. Azevedo, D. Morrow, J. Graumlich, A. Willemsen-Dunlap, M. Hasegawa-Johnson, T. Huang, K. Gu, S. Bhat, T. Sakakini, V. Sadauskas, and D. Halpin. Using conversational agents to explain medication instructions to older adults. In *Proceedings of the American Medical Informatics Association Fall Symposium*, 2018.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016. URL `https://arxiv.org/abs/1607.04606`.

A. Bordes, Y.-L. Boureau, and J. Weston. Learning End-to-End Goal-Oriented Dialog. *arXiv:1605.07683 [cs]*, May 2016. URL `http://arxiv.org/abs/1605.07683`.

F. Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015.

L. Dong, J. Mallinson, S. Reddy, and M. Lapata. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886. Association for Computational Linguistics, 2017. URL `http://aclweb.org/anthology/D17-1091`.

P. A. Duboue and J. Chu-Carroll. Answering the question you wish they had asked: The impact of paraphrasing for question answering. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 33–36. Association for Computational Linguistics, 2006.

J. Ganitkevitch, B. Van Durme, and C. Callison-Burch. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, 2013.

Google. Google Cloud Translation API Documentation | Translation API. https://cloud.google.com/translate/docs/.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. *Proceedings of NAACL-HLT 2018, New Orleans, Louisiana, June 1-6, 2018*, pages 1875–1885, 2018.

K. Jokinen and M. McTear. *Spoken dialogue systems*, volume 2 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers, 2009.

J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 282–289, 2001.

G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT 2016*, volume 5805, pages 260–270, San Diego, California, June 12-17, 2016, 2016. ACL.

L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Lau, et al. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 2018.

N. Madnani and B. J. Dorr. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387, 2010.

M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202.

G. Tur and R. D. Mori. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. John Wiley & Sons, May 2011. ISBN 978-1-119-99394-0. Google-Books-ID: RDLyT2FythgC.

Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, P. Kingsbury, and H. Liu. A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, 87: 12–20, Nov. 2018. ISSN 1532-0464. doi: 10.1016/j.jbi.2018.09.008. URL `http://www.sciencedirect.com/science/article/pii/S1532046418301825`.

J. Weston, A. Bordes, S. Chopra, A. M. Rush, B. van Merriënboer, A. Joulin, and T. Mikolov. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. *arXiv:1502.05698 [cs, stat]*, Feb. 2015. URL `http://arxiv.org/abs/1502.05698`.

S. Young, M. Gašić, B. Thomson, and J. D. Williams. POMDP-Based Statistical Spoken Dialog Systems: A Review. *Proceedings of the IEEE*, 101(5):1160–1179, May 2013. ISSN 0018-9219. doi: 10.1109/JPROC.2012.2225812.

W.-N. Zhang, Z.-Y. Ming, Y. Zhang, T. Liu, and T.-S. Chua. Exploring key concept paraphrasing based on pivot language translation for question retrieval. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 410–416. AAAI Press, 2015. ISBN 0-262-51129-0. URL `http://dl.acm.org/citation.cfm?id=2887007.2887065`.

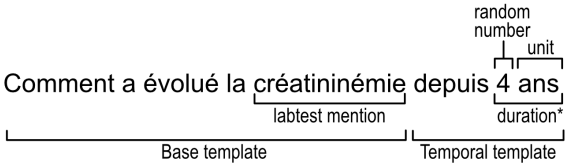
Supplementary material

	date	lab mention
mentions in the test set	34	177
Median length[min-max]	3 [0 - 6]	2 [1 - 11]
vocabulary in train set	1,364	451
vocabulary in test set(intersection with train)	58(0.38)	250(0.28)

Table S1: Description of the term mentions

	Utterances	Templates	Lab mentions	Words (*)	OOVs (*)	Perplexity (*)
training	16,000	170	336	144,850 (140,492)	-	-
development	4,000	53	73	36,211 (36,211)	4,724 (4,544)	137.5 (171.1)
test	178	-	-	1,579	467 (390)	194.5 (240.1)

Table S2: Characteristics of the datasets. (*) with paraphrases. OOV = out of vocabulary



Base template	170
Temporal templates	23
Labtest mentions	336
Duration units	15
Random duration range	[1-30]
Date range	[01/01/1950 - 31/12/2017]
Events	17

* duration can be replaced by date or event. The example means: 'How has creatininemia evolved in the last 4 years'

Figure S1: Combinations of templates and modifiers.

S1 Tuning parameters

For each model (except ELMo) we added some standard features to the input: normalized lemmas and part-of-speech (POS) tagging. Then, the sequence labelling part of the model was constituted of a CRF only or 2 layers of biLSTM or 2 layers of biLSTM followed by a CRF. The tuning parameters were: the dimension of the embeddings (50, 100, 300) except for ELMo (fixed to the default dimension), the number of units units in the biLSTM (64, 128, 256), the fraction of dropout after the embedding layer and after the LSTM layers (0.1, 0.2, 0.3, 0.4, 0.5). Regarding the classification part of the model, it constituted of a 1 dimensional convolution layer (2 to 5 filter kernel size and 50 to 250 filters, ReLu activation) followed by a max-pooling layer. Models were tuned using a random sample of parameters. The optimization function was ADAM. All the models were implemented using Keras [Chollet et al., 2015] with a Tensorflow [Abadi et al., 2015] backend.

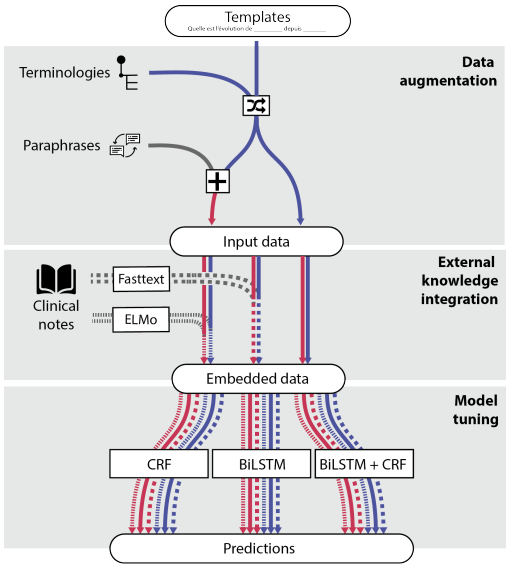


Figure S2: Experiences flow.

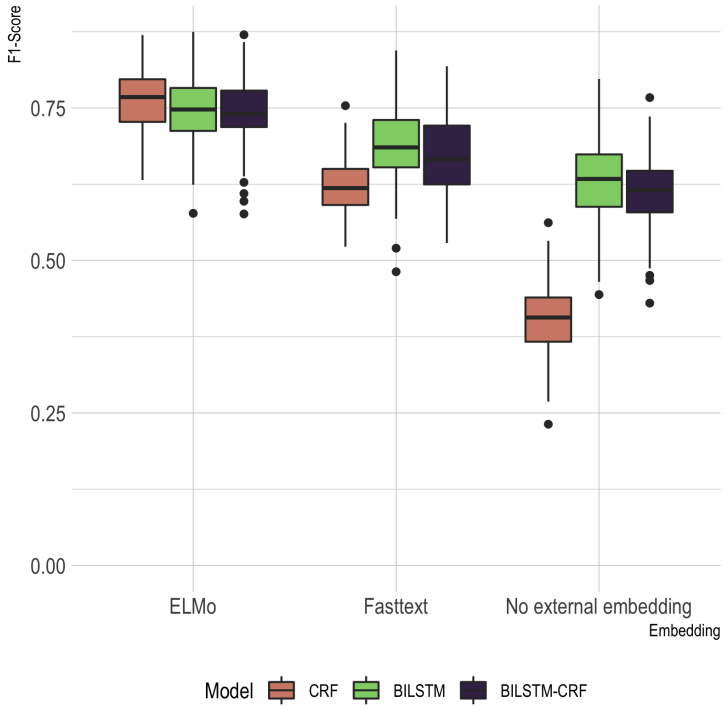


Figure S3: VA Task - sequence labelling

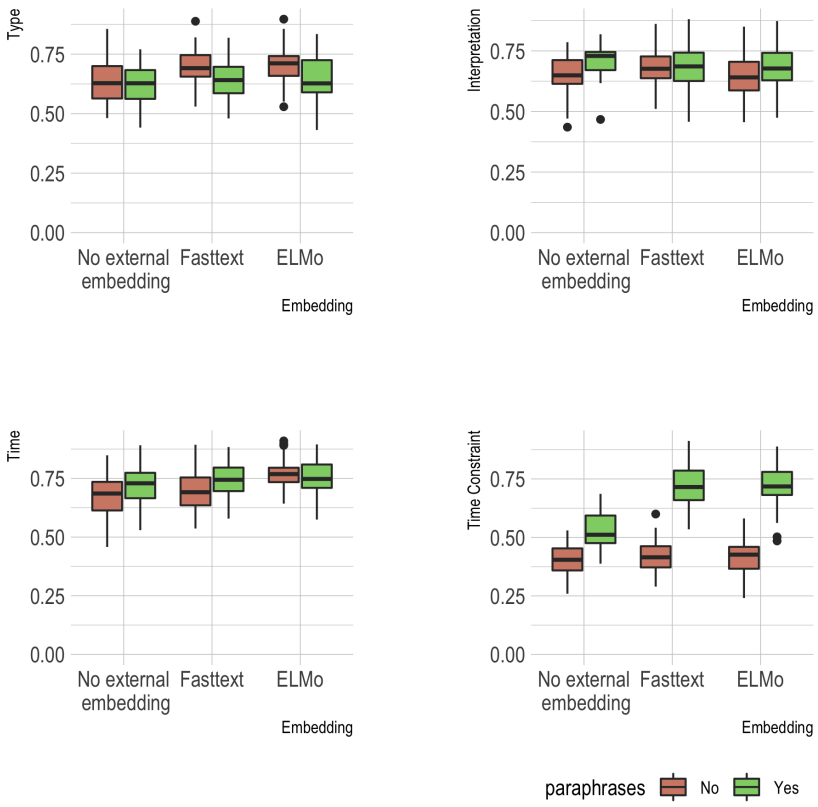


Figure S4: VA Task - classification

2.9 *Article : The Impact of Specialized Corpora for Word Embeddings in Natural Langage Understanding*

432 Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200197

The Impact of Specialized Corpora for Word Embeddings in Natural Langage Understanding

Antoine NEURAZ[a,b], Bastien RANCE[a], Nicolas GARCELON[a], Leonardo Campillos LLANOS[b], Anita BURGUN[a], Sophie ROSSET[b]
[a] *INSERM, UMR 1138 Team 22, Paris Descartes, Paris, France*
[b] *LIMSI, CNRS, Université Paris Saclay*

Abstract. Recent studies in the biomedical domain suggest that learning statistical word representations (static or contextualized word embeddings) on large corpora of specialized data improve the results on downstream natural language processing (NLP) tasks. In this paper, we explore the impact of the data source of word representations on a natural language understanding task. We compared embeddings learned with Fasttext (static embedding) and ELMo (contextualized embedding) representations, learned either on the general domain (Wikipedia) or on specialized data (electronic health records, EHR). The best results were obtained with ELMo representations learned on EHR data for the two sub-tasks (+7% and +4% of gain in F1-score). Moreover, ELMo representations were trained with only a fraction of the data used for Fasttext.

Keywords. Natural Language processing, Contextual word embeddings, Natural language understanding

1. Introduction

The recent advances in language representation such as static word embeddings [1,2], and contextual word embeddings (e.g. ELMo[3]) led to a significant performances improvement in natural language understanding (NLU). These methods can take advantage of large text corpora to learn a representation of the vocabulary in the vector space according to words semantics and syntactic use[1]. These approaches are unsupervised: they do not require learning datasets with manual annotations. However, the training corpus must be large enough to learn a correct representation. For example, it is customary to use a full dump of Wikipedia® to learn the models[1–3]. In the biomedical domain, the vocabulary is very specific. Moreover, when looking at clinical reports in electronic health records (EHRs), the syntax is also different from articles from sources similar to Wikipedia. A recent study from Wang *et al.*[4] evaluated the performances of embeddings learned on Wikipedia, biomedical publications and clinical notes in English : learning embeddings using clinical notes from EHRs increased the performances. Would this improvement also be true for contextual word embeddings? Nonetheless, it may be challenging to have access to the large number of documents requested (*e.g.* more than 100k patients in the study of Wang *et al.*) or even have access to embedding matrices learned on such corpus, due to privacy issues. It is possible that using more advanced methods to learn the word representations such as

ELMo compared to static embeddings would allow researchers to use embeddings learned on general domain corpora and still obtain good performances.

The aim of this work is to explore the impact of the method and data source of embeddings in the context of a weakly supervised NLU task in French. For this task in a restricted domain (biomedicine), we focused on the two NLU tasks: slot-filling and intent classification.

More specifically, this task aims at providing NLU in a virtual assistant in French for clinicians to explore biological tests results information in the patient's record in natural language (e.g. *Donne moi le dernier résultat de créatinine* 'Give me the last result of creatinin'). The set of queries that a physician may have about characteristics and results of a patient is broad and diverse. Therefore, enabling queries in natural language may help accessing information more efficiently. Given that no public dataset is available for this task in French, we used a weakly supervised approach by generating training data from question templates, terminologies and paraphrases as described in a previous work [5].

We compared the performances of two types of word representations: static word embeddings (Fasttext) and contextualized word embeddings (ELMo). For each method, we also compared two different learning sets in French: #1 Wikipedia or #2 a set of 1M clinical notes from our local EHR.

2. Methods

2.1. Embeddings

We compared two types of embeddings: #1 continuous skip-gram model with sub-word information as implemented in fastText[2], #2 embeddings from language models (ELMo) where the vectors are learned from the internal states of a deep bidirectional language model[3]. As a baseline, we use a continuous skip-gram model learned only on the training set (no external dataset).

We also compared the performances of these methods when learned on either a general domain dataset or a specialized dataset. The general domain dataset (hereafter, Wiki) is made of a dump of the French version of Wikipedia plus the French dataset of CommonCrawl. For this dataset, we used pre-learned models for French downloaded from fasttext website[1] for Fasttext and from github[2] for ELMo.

The specialized dataset (hereafter, EHR) is constituted of a random set of 1M clinical notes from the clinical data warehouse of Necker – Enfants malades hospital, a French AP-HP childrens hospital in Paris[6]. This dataset contains 162M tokens and a vocabulary of size 92k. For Fasttext, we used vectors of 300 dimensions and a window-size of 5. For ELMo, we kept only a subset of the EHR data (24M tokens) due to the high training time of ELMo. To compare embeddings from different sources, we kept the hyper-parameters as described in [3].

[1] https://fasttext.cc

[2] https://github.com/HIT-SCIR/ELMoForManyLangs

2.2. Task

We evaluated the impact of the different embeddings approaches on the task of NLU in a virtual assistant (VA task). Given that no public dataset is available for this task in French, we used the dataset generated from templates, terminologies and paraphrases as described in a previous work [5]. The training dataset contains 16,000 questions, 144k words (mean length of a question is 9), the development set of 4,000 questions for the tuning of the models. For the evaluation, we collected from physicians in our hospital, a set of 178 questions that they would like to ask in such a system. This set of questions has been manually annotated.

The slot filling task (VA-sequence) consists of a sequence labeling task aiming at identifying, in the question, the labtest mention (e.g *créatinine* 'creatinin') and the date- related information (e.g. *22/04/2012*, *depuis 4 semaines* 'for 4 weeks'). In the training set, the number of distinct lab mentions was 336 with a length ranging from 1 to 11 tokens and a median length of 2. In terms of vocabulary, there is only an overlap of 28% between the training set and the test set for labtest mentions. The date labels include exact dates, relative dates and time ranges.[5]

The intent classification task (VA-classification) is divided into 4 sub-tasks corresponding to 4 axes of classification. For each utterance, we assign one label per axis. Two axes concern the results of the lab exams (*i.e.* the type of result (5 categories) and interpretation of the result (5 categories). The two latter concern temporal aspects (*i.e.* the time of result (3 categories) and constraints on time (4 categories)).[5]

2.3. Models

We evaluated 5 different configurations of embeddings: continuous skip-gram model of 300 dimensions (D) learned on the training set; fasttext of 300D learned on Wiki; fasttext of 300D learned on EHR; ELMo of 1024D learned on Wiki; ELMo of 1024D learned on EHR.

For the sequence labeling task (VA-sequence), we used a recurrent neural network (RNN) based on bidirectional long short term memory units (biLSTM)[7]. We used two layers of biLSTM of size 256.

For the classification tasks we used a convolutional neural network (CNN)[8]. The model contains a 1D convolutional layer of 250 units, kernel size of 3, ReLU activation, followed by a max pooling layer and a dense fully connected layer.

All the models were implemented using Keras, with a Tensorflow backend, and the optimizer was Adam. We used dropouts after the embedding layer and before the final dense layer as well as L2-regularization on the convolutional layers to limit overfitting. We used a weighted F1-score (harmonic mean of the precision and the recall) to evaluate the results of the different models. To estimate the variability of the performances, we used 10 repetitions of 5 fold cross-validation on the test set.

3. Results

All the results are presented in Table 1. For the sequence labeling task (VA-sequence), the best results are obtained with ELMo learned on EHR with a F1-score of 0.76 (95%CI [0.74-77]) compared to Fasttext on EHR (0.67, 95%CI [0.61-0.73]). Interestingly, we

show similar results between ELMo on Wiki (0.69, 95%CI [0.67-0.70]) and Fasttext on EHR (0.67, 95%CI [0.61-0.73]). To note, the distance between the training set and the test set, estimated using the perplexity of the n-gram model, is 194.

Table 1. Results for the NLU task for the virtual assistant: sequence labeling (Bi-LSTM) and intent classification (CNN)

Method	**Sequence labeling** F1-score [95%CI]	**Intent classification** Mean F1-score [95%CI]
Baseline (only training set)	0.62 [0.61-0.64]	0.62 [0.58-0.65]
Fasttext on Wiki	0.69 [0.67-0.70]	0.52 [0.49-0.55]
Fasttext on EHR	0.67 [0.61-0.73]	0.66 [0.63-0.69]
ELMo on Wiki	0.69 [0.67-0.70]	0.49 [0.46-0.52]
ELMo on EHR	**0.76 [0.74-0.77]**	**0.70 [0.67-0.73]**

For the intent classification task (VA-classification), again, the best results come from the ELMo model learned on EHR (F1-score 0.70, 95%CI [0.67-0.73]) compared to the second-best Fasttext on EHR (F1-score 0.66, 95%CI [0.63-0.69]). In this task, the models learned on Wiki performed poorly: Fasttext on Wiki with a F1-score of 0.52 (95%CI [0.49-0.55]) and ELMo on Wiki 0.49 (95%CI [0.46-0.52]) with a baseline at 0.62 (95%CI [0.58-0.65]).

4. Discussion

The use of pre-learned embeddings may improve substancially the results on certain downstream tasks. As shown by Wang *et al.* [4] on various tasks in English, we show increased performances on two of three tasks in French when using pre-learned embeddings on a large corpus of clinical notes. The higher improvement was obtained on the VA-sequence task suggesting that tasks requiring extended lexicons may benefit the most of the pre-embeddings. Indeed, the identification of labtest mentions highly depends on the vobabulary available during training and we show that 72% of the mentions in the test set are absent from the training set. In this scenario, the representations learned on a large specialized dataset helped the model to obtain a better generalization when dealing with previously unseen mentions.

The test set shows 29% of out of vocabulary words overall. It may also partly explain the improvements observed on the classification task also (VA-classification). Interestingly, in the VA-classification task, both embeddings (Fasttext and ELMo) learned on Wiki did worsen the results compared to the baseline. This noise might come from the lack of specialized vocabulary in the Wiki corpus and a different usage of specialized words that may vary depending on the biomedical context and that are not taken into account in the general domain.

Sheikhshab *et al.*[9] evaluated ELMo embeddings on named entity recognition tasks in the biomedical domain. They also showed improved results when using an in-domain corpus to train ELMo (*i.e.* articles from Pubmed). It would be interesting to evaluate ELMo embeddings learned on a corpus of biomedical article in French but collecting such a corpus may be challenging.

This study has some limits. First, the embeddings learned on EHR came from a single hospital. This may cause some bias in the results given that the semantic particularities of the different practices may have an impact on the suitability of the embeddings for the different tasks. One possible solution to tackle this issue would be to

learn embeddings from multiple sites. However, due to privacy issues, it is not possible to transfer massive amount of hospital data. Two other approaches might be used to workaround this issue and should be explored for the biomedical domain: #1 the use of federated learning to train the models without moving the data[10]; #2 learning one embedding per site and aligning the different embeddings using unsupervised techniques. Finally, it is not obvious to determine the size of the training corpus for an embedding. It is a common practice to use all the articles from Wikipedia for example. But do we need that amount of data in a specialized domain? Comparing the impact of the size of the embedding corpus on downstream task will be of great interest.

5. Conclusions

Depending on the task, embeddings learned on large corpora can have a significant impact on NLP tasks in the biomedical domain in French. Moreover, learning these embeddings on clinical notes will increase the performances compared to general domain. As it may not be feasible to access a large corpus of clinical notes, it is still profitable to use advanced methods such as ELMo learned on general domain and obtain reasonable results. When the task does not rely on a large specialized vocabulary, the impact of external embeddings might be reduced.

References

[1] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean, Distributed Representations of Words and Phrases and their Compositionality, in: C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 26, Curran Associates, Inc., 2013: pp. 3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf (accessed March 31, 2017).

[2] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, Enriching word vectors with subword information, *ArXiv Preprint ArXiv:1607.04606*. (2016). https://arxiv.org/abs/1607.04606 (accessed February 14, 2017).

[3] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, Deep Contextualized Word Representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018: pp. 2227–2237. doi:10.18653/v1/N18-1202.

[4] Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, P. Kingsbury, and H. Liu, A comparison of word embeddings for the biomedical natural language processing, *Journal of Biomedical Informatics*. **87** (2018) 12–20. doi:10.1016/j.jbi.2018.09.008.

[5] A. Neuraz, A. Burgun, L.C. Llanos, and S. Rosset, Natural language understanding for task oriented dialog in the biomedical domain in a low ressources context, *Machine Learning for Health (ML4H) Workshop at NeurIPS 2018*. (2018).

[6] N. Garcelon, A. Neuraz, R. Salomon, H. Faour, V. Benoit, A. Delapalme, A. Munnich, A. Burgun, and B. Rance, A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse, *Journal of Biomedical Informatics*. **80** (2018) 52–63. doi:10.1016/j.jbi.2018.02.019.

[7] A. Graves, and J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, *Neural Networks*. **18** (2005) 602–610. doi:10.1016/j.neunet.2005.06.042.

[8] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, *Proceedings of the IEEE*. **86** (1998) 2278–2324. doi:10.1109/5.726791.

[9] G. Sheikhshabbafghi, I. Birol, and A. Sarkar, In-domain Context-aware Token Embeddings Improve Biomedical Named Entity Recognition, (n.d.) 5.

[10] D. Zhang, X. Chen, D. Wang, and J. Shi, A Survey on Collaborative Deep Learning and Privacy-Preserving, in: 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC), 2018: pp. 652–658. doi:10.1109/DSC.2018.00104.

Chapitre 3

APMed : un corpus de textes cliniques en français annotés pour les informations sur les médicaments

Comme nous avons pu le constater dans le chapitre introductif, construire des modèles d'extraction des informations sur les médicaments est un enjeu important en terme de santé publique. En langue française, il n'existe pas de système qui permette d'effectuer cette tâche avec des performances voisines de l'état de l'art. Le premier frein au développement d'un tel système est l'absence de jeu de données permettant d'entraîner des modèles de machine learning performants. Notre premier objectif ici a donc été de proposer un tel jeu de données.

Dans le chapitre précédent, nous avons également mis en évidence que l'injection de connaissances (latentes ou expertes) pouvait permettre d'améliorer les performances de modèles appris sur les données. Nous nous sommes donc attaché à construire un modèle d'extraction hybride intégrant règles expertes et réseaux de neuronnes récurrents, ainsi qu'un modèle séquentiel opérant une extraction jointe des entités et de leurs relations.

3.1 Matériels et méthodes

3.1.1 Extraction des données

Les données proviennent de l'entrepôt de données de santé l'APHP (EDS).[170] Cet entrepôt regroupe les données produites pendant le soin (e.g. compte-rendus médicaux, résultats d'examens de biologie, codes diagnostics, codes d'actes) dans les 39 hôpitaux de la région parisienne qui font partie de l'APHP. L'EDS est organisé dans un format d'entrepôt en étoile (format i2b2).[171]

A partir de l'EDS, nous avons extrait 1 million de documents de manière aléatoire sur l'ensemble des documents textuels disponibles en avril 2018. Les documents ont été choisis parmi les types de documents suivants, susceptibles de contenir des informations sur les traitements des patients : compte-rendu (CR) Bilan Post-Opératoire,CR Bilan Pré-Opératoire,CR Consultation,CR Coronarographie,CR Nutrition, CR Nutrition, Consultation Initiale, CR Passage Urgences, compte-rendu d'hospitalisation (CRH) Chirurgie, CRH Evolution,CRH Hématologie,

CRH Hopital, CRH de Jour, Lettre Sortie (CRH provisoire), CRH Néonatologie, CRH Neurologie, CRH Neuropathies périphériques, CRH Orthopédie, CRH Pédiatrie générale, CRH Service, CR Staff, Lettre Autre, Lettre Consultation, Lettre Consultation Initiale, Lettre Consultation Suivi, Lettre CRH, Lettre CR tout venant, Ordonnance, Lettre Sortie, Réunion de Concertation Pluridisciplinaire.

L'ensemble des documents ont été utilisés pour entraîner ou fine-tuner différents types d'embeddings utilisés dans des expériences décrites plus loin : word2vec, fasttext, ELMo.

Parmi les documents extraits, 320 ont été sélectionnés aléatoirement pour être annotés manuellement par trois médecins selon les règles décrites dans la section suivante.

3.1.2 Annotation

Nous avons utilisé comme base pour développer notre guide d'annotation celui de challenge i2b2 2009 sur l'extraction d'information relatives aux médicaments.[139] En effet, au moment de la réalisation de ce travail, les jeux de données plus récents évoqués dans les sections précédentes n'étaient pas encore disponibles.

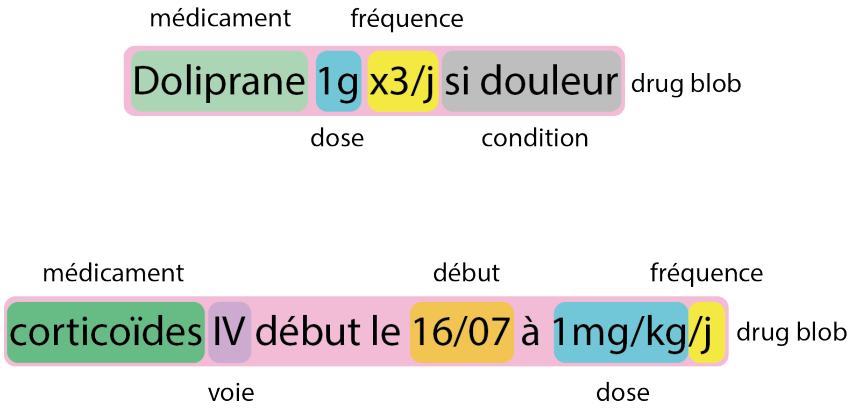


FIGURE 3.1 : Deux exemples d'annotations

Nous avons annoté des entités, des évènements et des attributs (Figure 3.1) :

- 7 types d'entités : nom du médicament et classe ATC, posologie, fréquence, durée, voie d'administration et condition de la prise). (Table 3.1)
- 6 types d'évènements : début, fin, changement ou prolongement de médicament, diminution ou augmentation de la quantité (soit par sa posologie, soit par sa fréquence) (Table 3.2)
- 3 types d'attributs : temporalité (prescription ou événement du passé, présent ou futur), certitude (factuel, incertain, suggéré, conditionnel, négation) et sujet de la prescription (patient, famille, autre) (Table 3.3)

Pour chaque nom de médicament ou de classe médicamenteuse repérés, les champs de prescription associés sont annotés (e.g. posologie, fréquence). Le cas échéant, des évènements s'y rapportant sont également annotés (e.g. début, fin). Pour chaque nom de médicament identifié, les attributs de temporalité, certitude et sujet sont ajoutés l'annotation de l'entité.

Enfin, pour représenter les relations entre les mentions de médicaments et les entités/évènements s'y rapportant, nous avons choisi d'ajouter une méta-classe d'entité représentant schématiquement une ligne de prescription (*drug blob*). Autrement dit, le span contenant une mention de médicament

Type	Description	Exemple
Nom du médicament	médicament, molécule active, association ou protocole	doliprane, paracétamol, Augmentin
Classe de médicament	classe ATC ou thérapie courante	ß-Lactamine, antibiothérapie
Dosage	Dose ou concentration	3 mg, 2 comprimés
Fréquence	Fréquence d'administration	3 par jour, tous les matins
Durée	Durée de la prescription	3 semaines, jusqu'à l'opération
Voie	Mode d'administration	intraveineuse, per os
Condition	L'évènement qui provoque l'administration	si douleur, si infection

TABLE 3.1 : Description des entités annotées

Type	Description	Exemple
Start	date ou mot signalant le début de l'utilisation	antibiotherapie debutee lors de la chirurgie
Stop	date ou mot signalant l'arrêt de l'utilisation	a arrete a j5
Start-Stop	prise unique	hospitalisation de jour du 27/12/2012 pour sa perfusion
Increase	augemtation de la dose ou de la fréquence	augmentation des doses de morphine
Decrease	diminution de la dose ou de la fréquence	diminution des doses de morphine
Continue	maintien sans changement	Pas de modification de la corticothérapie

TABLE 3.2 : Description des événements affectant les entités annotées

et toutes les informations s'y rapportant sont encapsulés dans une grande entité *drug blob*. Le guide d'annotations complet est accessible en ligne à l'adresse : `https://equipe22.github.io/medExtAnnotation`. Il est également disponible dans l'Annexe A.

Ce corpus est accessible sous réserve d'acceptation du comité scientifique et éthique de l'APHP (https://recherche.aphp.fr/eds/recherche/).. La présente étude a été autorisée par le comité scientifique éthique sous le numéro (CSE-2018-25).

3.1.3 Développement d'un système hybride d'extraction des informations sur les médicaments (Première Phase)

Le système d'annotation hybride est constitué de 2 composants (Figure 3.2) :

- un système à base de dictionnaires et de règles expertes est utilisé pour pré-annoter les documents
- un modèle de type BiLSTM-CRF associé à des word embeddings ELMo utilise comme entrée le texte et la pré-annotation

Type	Description	Exemple
Temporalité		
Past	date ou mot signalant le début de l'utilisation	le patient a pris du paracetamol du 25/12 au 27/12
Present	date ou mot signalant l'arrêt de l'utilisation	patient sous paracetamol
futur	prise unique	3e injection de paracetamol prévue le 5/01
Certitude		
Conditional	sous certaines conditions	paracetamol si fièvre
Suggestion	suggéré mais indépendant d'une condition	nous envisageons de mettre le patient sous paracetamol
Factual	certain (par défaut)	paracetamol ce jour
Negated	négatif	pas de paracetamol ce jour
Contraindicated	contre-indication	allergie au paracetamol

TABLE 3.3 : Description des attributs modifiant les entités annotées

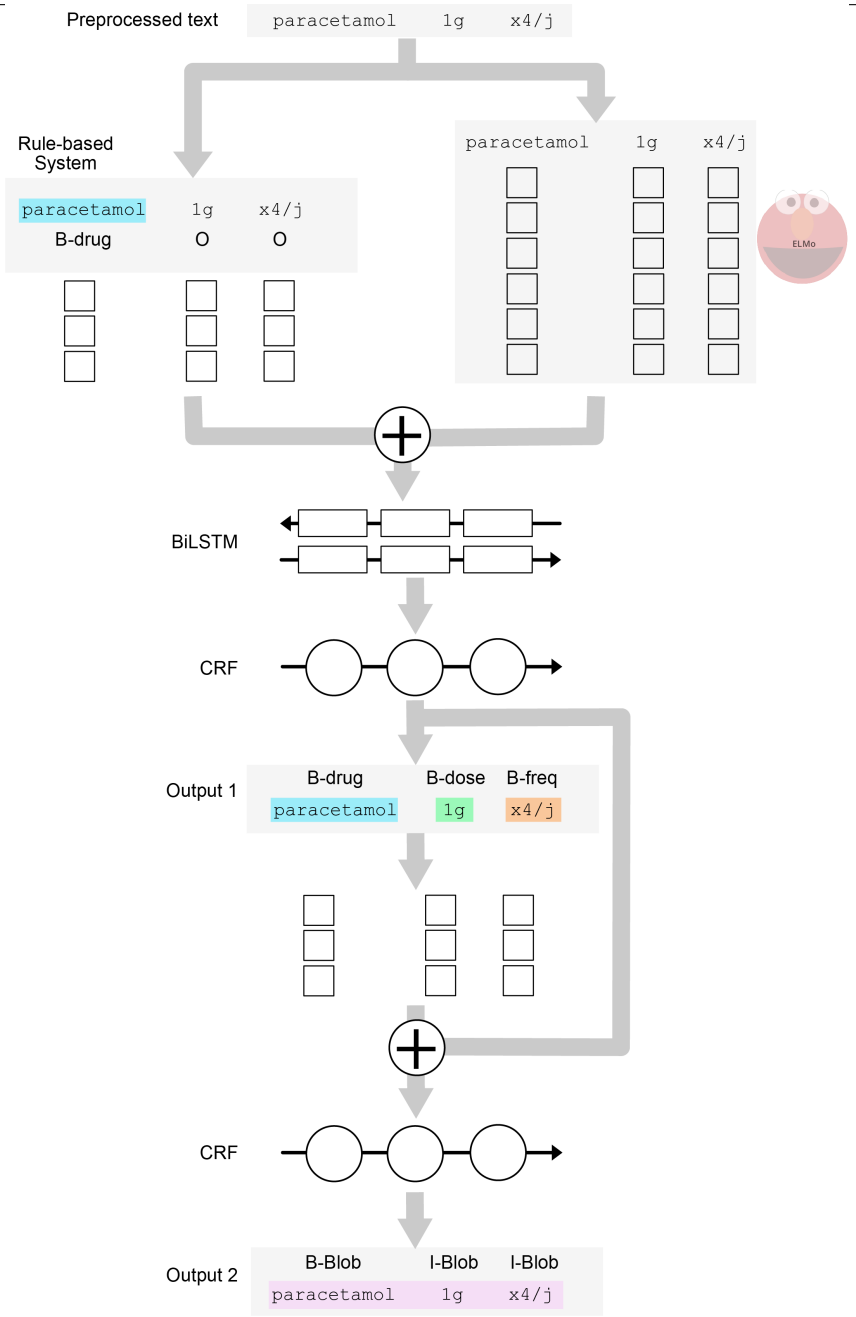


FIGURE 3.2 : Schéma du modèle hybride d'annotation

Le système à base de règles

La première étape du système permet de repérer les noms de médicaments et de classes par correspondance exacte (*exact match*) à partir d'un dictionnaire. Ce dictionnaire a été constitué à l'aide d'une combinaison de dictionnaires disponibles en ligne : la base de données publique des médicaments de l'agence nationale de sécurité des médicaments [1] et la base de données des médicaments remboursés par l'assurance maladie (OpenMedic) [2] . Nous avons ainsi obtenu un

1. http://base-donnees-publique.medicaments.gouv.fr
2. https://www.data.gouv.fr/en/datasets/open-medic-base-complete-sur-les-depenses-de-medicaments-interregimes/

dictionnaire contenant 9557 mentions de médicaments .

La détection d'une mention de médicament a ensuite servi d'ancre pour une série d'expressions régulières visant à extraire les autres entités.

Les résultats du système à base de règles sont ensuite converties au format BIO [172] afin d'être passés au modèle d'apprentissage.

Le modèle d'apprentissage

Nous avons comparé les performances de plusieurs types d'embeddings : word2vec sur les données d'apprentissage, Fasttext sur les données d'EHR, et ELMo sur les données d'EHR. Les embeddings Fasttext (modèles skip-gram avec informations sur les ngrams de caractères) ont été entraînés sur 1M de documents (dimension 300, fenêtre de 5 tokens). Les embeddings contextuels ELMo ont été entraînés sur un échantillon de 100k documents non annotés issus de l'EDS. Nous avons utilisé un nombre plus faible de documents à cause de contraintes techniques. Pour cela nous avons utilisé la librairie `bilm-tf` d'AllenAI. [3] Nous avons conservé les paramètres par défaut pour l'entraînement du modèle ELMo (dimension 1024).

La partie BiLSTM-CRF est constituée de 2 couches de BiLSTM suivi d'un CRF (CRF_{entits}) pour la prédiction des tokens d'entités. Un deuxième CRF ($CRF_{drug\ blob}$) prenait en entrée la sortie du BiLSTM concaténée avec la sortie du CRF_{entits}. Les modèles ont été entraînés pour un maximum de 50 epochs avec un optimiseur de type ADAM et un *learning rate* de 0.001 associé à un *early stopping* (patience = 8 epochs). Les modèles ont été implémentés avec la library keras et un backend tensorFlow. Les hyperparamètres ont été optimisés par recherche aléatoire sur 15 itérations à l'aide de la librarie Hyperas.

3.1.4 Développement d'un système séquentiel d'annotation (Deuxième phase)

La deuxième phase a été développée ultérieurement, avec pour vocation d'être déployée à grande échelle. Nous n'avons pas évalué de système hybride dans cette configuration (pas de système à base de règles expertes) mais un système séquentiel composé de deux modèles connectés. Nous avons utilisé les embeddings de type BERT multilingue [34] que nous avons fine-tuné sur un ensemble de 10 millions de documents. Les embeddings BERT étaient constitués de la concaténation des 4 dernières couches de BERT. Nous avons pour cela utilisé la librairie FLAIR [4] [173] et son implémentation du BiLSTM-CRF.(Figure 3.3) Deux modèles distincts ont été entraînés : le premier pour les entités et le second pour les relations (représentées sous forme de méta-entités *drug-blob*). Les annotations issues du premier modèle été concaténées sous forme d'embeddings de dimension 10 aux entrées BERT du second modèle. Les modèles utilisés étaient des BiLSTM à 2 couches avec un décodeur CRF, avec 1024 unités par couche. L'optimisation était faite par *asynchronous stochastic gradient descent* (ASGD) [174] avec réduction du taux d'apprentissage en cas de plateau.

3. https://github.com/allenai/bilm-tf
4. https://github.com/flairNLP/flair

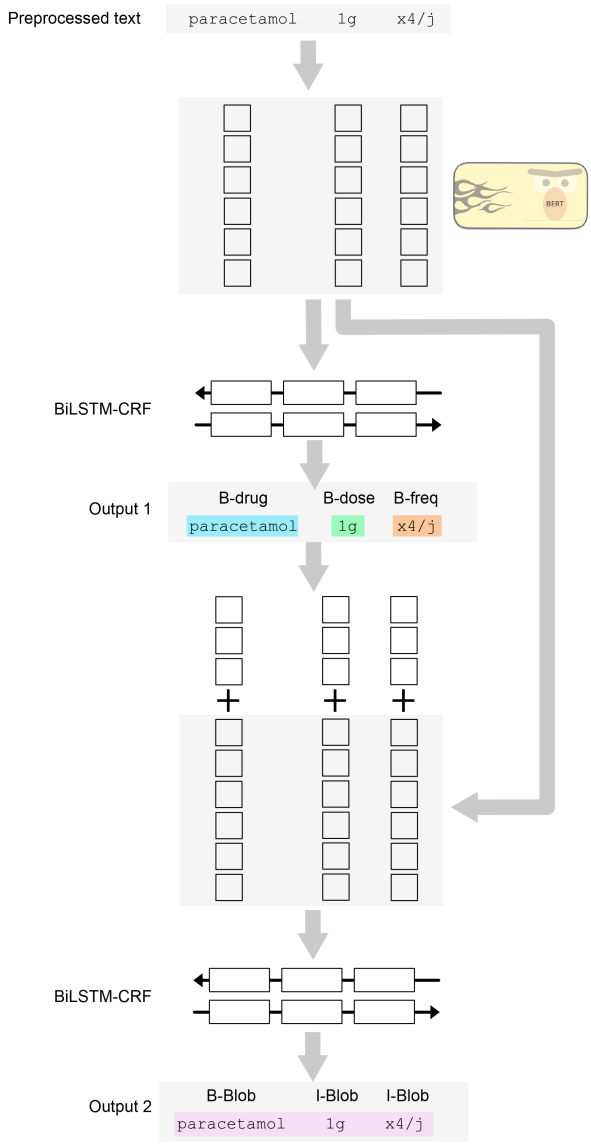


FIGURE 3.3 : Schéma du modèle d'annotation basé sur BERT

Métriques d'évaluation

Voici les métriques d'évaluation que nous avons utilisées :

$$Rapple = \frac{VP}{VP + FN}$$

$$Prcision = \frac{VP}{VP + FP}$$

Donnée	Corpus entier[a]	Entrainement[a]	Développement[a]	Test
Entités				
Nom de médicament	1768 (1,12)	1227 (1,13)	143 (1,02)	398 (1,13)
Classe ATC	334 (1,33)	228 (1,36)	30 (1,27)	76 (1,28)
Dosage	1134 (1,84)	761 (1,8)	62 (1,85)	311 (1,95)
Fréquence	830 (2,67)	600 (2,67)	46 (3,09)	184 (2,54)
Durée	120 (2,13)	70 (2,3)	13 (2)	37 (1,84)
Voie d'administration	148 (1,16)	85 (1,12)	8 (1)	55 (1,25)
Condition	92 (3,15)	61 (3,15)	3 (3)	28 (3,18)
Evènements				
Début	205 (2,62)	139 (2,63)	15 (2,87)	51 (2,51)
Arrêt	120 (3,42)	86 (3,42)	4 (2,75)	30 (3,5)
Prise unique	136 (3,46)	92 (3,52)	14 (4,14)	30 (2,93)
Switch	38 (1,45)	22 (1,59)	4 (1)	12 (1,33)
Prolongement	113 (1,42)	83 (1,39)	10 (1,9)	20 (1,3)
Augmentation	21 (1)	14 (1)	0(0)	7 (1)
Diminution	23 (2,3)	17 (2,53)	3 (1)	3 (2,33)
Relations				
drug blob	1275 (7,93)	885 (7,73)	79 (7,62)	311 (8,61)

[a] Nombre total d'entités (nombre moyen de tokens par entité)

TABLE 3.4 : Description des annotations dans le corpus APMed

$$F1score = \frac{Recall * Precision}{Recall + Precision}$$

avec VP = vrai positif, FN= faux négatif, FP = faux positif

3.2 Résultats

3.2.1 Description du corpus

Les 320 documents annotés représentaient 19 957 phrases et 173 796 tokens. Nous avons divisé le jeu de données en trois parties : 216 documents pour l'entraînement, 24 pour le développement et 80 pour le test. Cela représentait respectivement 13 737, 1373 et 4847 phrases. Une description du nombre de tokens et d'entités en fonction des classes d'annotation est disponible dans la Table 3.4. Au total, 5082 entités et évènements (1,8 tokens par entité en moyenne) ont été annotés ainsi que 1275 méta-entités *drug blob* (7,93 tokens en moyenne). Il est intéressant de noter qu'une méta-entité regroupe plusieurs relations entre une mention de médicament et les attributs de prescription. En effet, un *drug blob* pourra contenir les relations médicament-dose, médicament-fréquence, médicament-voie, médicament-durée,médicament-condition ainsi que toute relation médicament-événement. De plus chacune de ces relations peut-être représentée plusieurs fois en sein d'un même *drug blob*. Prenons comme exemple, le *drug blob* suivant : "*paracetamol 3cp/j puis 1cp/j*," il y a deux instances de dose (*3cp* et *1cp*) et de fréquence (*/j*, */j*) séparées par un évènement *diminution*.

Modèle	F-mesure	Précision	Rappel
Entités			
Règles	79.41	**94.67**	72.28
word2vec	73.93	83.89	67.57
Fasttext	88.08	89.48	87.17
ELMo	88.66	87.95	89.44
Règles + word2vec	83.74	88.46	80.24
Règles + Fasttext	88.18	91.73	85.54
Règles + ELMo	89.86	90.83	89.17
BERT*	**90 [88.5-91.4]**	89.2 [87.4-91.0]	**90.7 [88.8-92.4]**
Relations			
ELMo	49.87	45.52	55.29
BERT*	**92.6 [91.2-93.9]**	**89.3 [87.1-91.4]**	**96.1 [94.7-97.5]**

* Deuxième phase : système séquentiel

TABLE 3.5 : Résultats moyens des modèles BiLSTM-CRF sur le corpus APMed

3.2.2 Résultats des systèmes d'extraction d'information

Les résultats sont résumés dans les Tables 3.5 et 3.6.

Système hybride (Première phase)

De manière globale, les meilleurs résultats ont été obtenus avec le modèle hybride Règles+ELMo avec une F-mesure à 89.86. Cette performance s'explique principalement par l'augmentation du Rappel à 89.17 pour Règles+ELMo. En revanche la meilleure précision a été obtenue par le système à base de règles avec 94.67, contre 91.73 pour le meilleur modèle d'apprentissage (Règles+Fasttext). L'ajout d'embeddings appris sur les données médicales a permis un gain de performances notables, +14 et +15 points de F-mesure avec Fasttext et ELMo, respectivement, par rapport au modèle word2vec. Le gain de performances s'observait à la fois sur la précision et le rappel.

La combinaison avec les règles expertes a permis un gain plus modéré de performances globales sur les modèles entraînés sur données médicales : +0.1 et +1.2 de F-mesure pour Fasttext et ELMo. En fait, nous avons pu observer que la précision avait tendance à augmenter : +2.73 et +3.83 pour Fasttext et ELMo alors que le rappel avait tendance à diminuer -1.63 et -0.27. Le modèle word2vec, non entraîné sur les données médicales bénéficie de l'apport des règles +9.81 de F-mesure, +4.57 de précision et +12.67 de rappel.

Concernant les résultats sur les différentes classes, nous avons observé des résultats contrastés. (Table 3.6) La meilleure précision était quasiment systématiquement obtenue par le système à base de règles, sauf pour la durée où la précision des règles (49.25) était très faible. En revanche, le rappel était systématiquement meilleur avec les systèmes d'apprentissage. La combinaison Règle+ELMo a obtenu les meilleures F-mesures pour le nom du médicament (95.33), la classe ATC (64.36), le dosage (95.29), et la condition de prise (62.16). Pour la durée, la fréquence et la voie d'administration, les résultats de F-mesure étaient meilleurs sans les règles, 92.8, 82.17 et 75.52, respectivement.

	Règles			ELMo-BiLSTM-CRF			Règles + ELMo-BiLSTM-CRF			BERT-BiLSTM-CRF		
Label	F[a]	P[b]	R[d]	F	P	R	F	P	R	F	P	R
Entités												
Nom de médicament	90.31	**96.46**	84.89	92.2	93.79	90.67	**95.33**	95.33	95.33	93.8	92.1	**95.6**
Classe ATC	13.33	**87.5**	7.22	62.3	66.28	58.76	64.36	61.9	**67.01**	**69.1**	73.9	65.1
Dosage	90.43	**96.62**	84.98	92.17	91.13	93.23	**95.29**	95.52	**95.05**	93.4	93.0	93.9
Fréquence	86.13	**98.89**	76.28	92.8	93.3	92.31	92.24	93.04	91.45	**93.4**	94.2	**92.6**
Durée	48.89	49.25	48.53	82.17	86.89	77.94	78.79	81.25	76.47	**92.0**	**92.2**	**92.3**
Voie d'administration	47.92	**85.19**	33.33	**75.52**	72.97	78.26	72.86	71.83	73.91	72.2	62.5	**86.3**
Condition	33.64	**100**	20.22	55.9	62.5	50.56	**62.16**	77.97	51.69	55.5	54.4	**58.9**
Relations												
Drug Blob	-	-	-	49.87	45.52	55.29	-	-	-	**92.6**	**89.3**	**96.1**

[a] F1-score [b] Précision [c] Rappel

TABLE 3.6 : Résultat du système d'annotation hybride sur le corpus APMed

Système séquentiel (deuxième phase)

Ce système, basé sur des embeddings de type BERT, a obtenu une meilleure F-mesure (90 [88.5-91.4]) que le meilleur des modèles hybrides (89.86) (Règles+ELMo) alors même qu'il n'a pas été entraîné avec des pré-annotations par règles expertes. L'amélioration était marginale et probablement dans la marge d'erreur. Cette légère amélioration était principalement médiée par une augmentation du rappel 90.7 [88.8-92.4] contre 89.7 pour Règles+ELMo. En revanche, la précision était plus faible que le système de Règles 89.2 [87.4-91.0] contre 94.67, respectivement. Et même, moins bon que Règle+Fasttext (91.73)

Au niveau des relations (*drug blob*), l'amélioration par rapport au système hybride était significative : 92.6 de F-mesure contre 49.87.

3.3 Discussion

Nous proposons ici APmed, un jeu de données annotées pour les informations relatives aux médicaments sur des textes médicaux en français issus d'un entrepôt de données cliniques. Il s'agit, à notre connaissance, du premier jeu de données de ce type, disponibles pour la recherche (sous réserve de l'obtention d'un accord éthique et d'un conventionnement avec l'APHP). Les textes médicaux inclus proviennent de 39 hôpitaux différents sans sélection de services particuliers. Les jeux de données i2b2 2009 et n2c2, provenant de la base de données MIMIC, sont issus de services de réanimation uniquement. De plus, ils sont hétérogènes de part leur typologie (e.g., lettres, comptes-rendus, ordonnances). Ces éléments en font un jeu de données très proche de la "vie réelle." Ainsi, les modèles entraînés sur ce jeu de données pourront bénéficier de cette hétérogénéité et avoir de bonnes propriétés de généralisation. Nous proposons également un guide d'annotation, disponible librement, afin de permettre à la communauté de reproduire et d'étendre le travail exposé ici.

Ce jeu de données a des limites. Tout d'abord nous pouvons constater qu'il est de petite taille si l'on compare aux corpus disponibles en langue anglaise. A peine 10% de jeux de données comme n2c2 2018 par exemple, en terme de nombre d'entités annotés. D'autre part, bien que l'annotation ait été faite en concertation entre les experts sur les cas posant question, nous n'avons pas pu effectué de double annotation pour évaluer l'agrément inter-annotateur. La raison principale justifiant ces limite réside dans la quantité de ressources disponibles en terme de

temps médical pour l'annotation. Il faudra renforcer la qualité et la taille du corpus par une nouvelle campagne d'annotation quand de nouvelles ressources seront disponibles.

Notre dernière contribution consiste à proposer trois modèles :
- un modèle base-line basé sur un BiLSTM-CRF et des embeddings ELMo ;
- un modèle hybride combinant le précédent modèle avec des règles expertes ;
- un modèle séquentiel permettant l'extraction jointe des entités et des relations (sous forme de méta-entités), combinant des embeddings basés sur BERT et deux BiLSTM-CRFs.

Bien qu'il soit délicat de comparer des résultats qui portent sur des jeux de données différents et des modèles différents, les résultats que nous obtenons sont comparables avec ceux relevés sur les jeux de données du domaine. En effet, concernant les noms de médicament, les meilleurs résultats sur n2c2 sont à 95.6 de F-mesure [125], et 94.6 sur i2b2 2009 [123] comparés au 95.3 que nous rapportons. De la même façon, nous obtenons 93.4 de F-mesure sur l'extraction de la fréquence contre 97.5 sur n2c2 [125] et 92.4 sur i2b2 [131]. De même pour la dose, 95.2 sur APmed contre 94.8 sur n2c2 [125] et 93.0 sur i2b2 [123]. En revanche, les résultats de nos modèles sont plus faibles concernant la voie d'administration : 75.5 pour notre meilleur modèle contre 96.9 sur i2b2 et 95.6 sur n2c2. Il apparaît que dans les jeux de données anglo-saxons, l'expression de la voie d'administration est beaucoup plus formalisée que dans les textes français. La variabilité étant plus faible, les résultats des modèles d'extraction sont meilleurs.

S'agissant de l'extraction des entités, nous n'avons pas pu comparer les résultats obtenus ici avec ceux des autres jeux de données car notre approche considère d'un bloc l'ensemble des relations rattachées à une mention. Une option de comparaison pourrait être de considérer que la F-mesure que nous obtenons pour les *drug blob* est une moyenne de la F-mesure sur l'ensemble des classes. Avec cet assomption, nous obtenons une performance en retrait par rapport à l'état de l'art sur les autres jeux de données : 92.6 contre 97.2 sur n2c2 [125] et 96.2 sur i2b2 [135].

Il est intéressant de noter que les résultats que nous obtenons sont basés sur un jeu de données d'entraînement beaucoup plus petit. L'entraînement, ou le fine-tuning des embeddings contextuels de type ELMo et BERT sur de grands volumes de données cliniques en français a joué un rôle majeur dans l'amélioration des performances. Il existe encore une marge de progression relativement aux embeddings. En effet, ELMo avait été entraîné sur seulement 100 000 textes issus d'un unique hôpital. Et le modèle BERT que nous avons fine-tuné était un modèle multilingue. Il est probable que l'utilisation d'un modèle entraîné sur des textes français, comme CamemBERT [39], pourrait faire encore progresser les résultats.

D'autre part, l'injection d'information à travers les règles expertes améliore les performance générales du modèle à condition que la qualité de l'information injectée soit suffisante. En effet, sur l'extraction de la durée, le modèle avec les règles obtiens des performances plus faible que le modèle sans les règles. Hors, les performances des règles seules sur cette catégories sont très faibles (F-mesure à 48.9). Il semble que l'injection de connaissances externes ne soit bénéfique que si cette information ne contient pas trop bruit. D'autres voies d'intégration de connaissances pourront être bénéfiques à explorer comme notamment l'injection de connaissances ontologiques sur les relations entre les termes médicaux contenus dans les textes. Ces informations étant connues a priori, l'utilisation d'embeddings spécifiques des terminologies médicales, en addition aux modèles de langage comme ELMo ou BERT pourraient amener à une amélioration des performances des modèles. Les embeddings de Poincaré par exemple ont montré une bonne capacité à représenter dans des dimensions réduites les relations hiérarchiques de terminologies médicales.[175]

En conclusion, nous avons développé un jeu de données spécifique en langue française, annoté

pour les informations sur les médicaments (entités, évènements, attributs, relations), ainsi que des modèles d'extraction de ces informations. Les performances des modèles présentés permettent d'envisager leur utilisation à grande échelle au sein d'entrepôt de données de santé.

Cependant, afin de déployer ces modèles à grande échelle, il faut pouvoir les faire intéragir avec l'éco-système des données de santé, et produire des informations qui puissent être tracées jusqu'à leur origine. Pour aider à réaliser cette tâche, nous avons développé une trousse à outils permettant de faciliter l'intéraction de systèmes d'annotations diverses avec les données données de santé. Cet outil sera détaillé dans le Chapitre 4

La première phase de ce travail (sans les modèles BERT) fait l'objet d'un article actuellement en révision majeures dans le journal JMIR Medical Informatics. Il s'agit d'un travail collaboratif que j'ai encadré. Le preprint est accessible ici :

Jouffroy, Jordan, Sarah F Feldman, Ivan Lerner, Bastien Rance, Anita Burgun, et **Antoine Neuraz**.
« MedExt : Combining Expert Knowledge and Deep Learning for Medication Extraction from French Clinical Texts (Preprint) ».
Preprint.Journal of Medical Internet Research, 23 janvier 2020.
<https://doi.org/10.2196/preprints.17934>.

3.4 *Article : MedExt : combining expert knowledge and deep learning for medication extraction from French clinical texts*

MedExt: combining expert knowledge and deep learning for medication extraction from French clinical texts

Jordan Jouffroy[1, 2] MD; Sarah F Feldman[1] MD; Ivan Lerner[1, 2] MD; Bastien Rance[2, 3] PhD; Anita Burgun[1, 2] MD, PhD; Antoine Neuraz[2] MD

[1]Department of biomedical informatics Necker-Enfants malades Hospital APHP Paris FR
[2]INSERM UMRS 1138 team 22, Université de Paris Paris FR
[3]Department of biomedical informatics Georges Pompidou European Hospital APHP Paris FR

Corresponding Author:
Antoine Neuraz MD

Abstract

Background: Information related to patient medication is crucial for health care. However, up to 80% of the information resides solely in unstructured text. Manual extraction may be difficult and time-consuming. Many studies have shown the interest of natural language processing for this task but only a few on French corpus.

Objective: We aim at developing a system to extract medication-related information from French clinical text.

Methods: We developed a hybrid system combining an expert rule-based system (RBS), contextual word embedding (ELMo) trained on clinical notes and a deep recurrent neural network (BiLSTM-CRF). The task consists in extracting drug mentions and their related information (e.g. dosage, frequency, duration, route, condition). We manually annotated 320 clinical notes extracted from a French clinical data warehouse, to train and evaluate the model. We compared the performances of our approach to standard approaches: rule-based or machine learning only, and classic word embeddings. We evaluated the models using token level recall, precision and F-measure.

Results: Models including RBS, ELMo and BiLSTM reached the best results: overall F-measure of 89.9%. F-measures per category were 95.3% for the medication name, 64.4% for the drug class mentions, 95.3% for the dosage, 92.2% for the frequency, 78.8% for the duration, and 62.2% for the condition of the intake.

Conclusions: Associating expert rules, deep contextualized embedding (ELMo) and deep neural networks improves medication information extraction. Our results reveal a synergy when associating expert knowledge and latent knowledge.

(JMIR Preprints 23/01/2020:17934)
DOI: https://doi.org/10.2196/preprints.17934

Original Manuscript

Original Paper

Jordan Jouffroy, MD[1,2], Sarah F Feldman, MD[1,2], Ivan Lerner, MD[1,2], Bastien Rance[2,3], PhD, Anita Burgun, MD[1,2], PhD, Antoine Neuraz, MD[1,2]

[1] Department of biomedical informatics, Necker-Enfants malades Hospital, APHP, Paris, France
[2] INSERM UMRS 1138 team 22, Université de Paris
[3] Department of biomedical informatics, Georges Pompidou European Hospital, APHP, Paris, France

MedExt: combining expert knowledge and deep learning for medication extraction from French clinical texts

Abstract

Background: Information related to patient medication is crucial for health care. However, up to 80% of the information resides solely in unstructured text. Manual extraction may be difficult and time-consuming. Many studies have shown the interest of natural language processing for this task but only a few on French corpus.

Objective: We aim at developing a system to extract medication-related information from French clinical text.
Methods: We developed a hybrid system combining an expert rule-based system (RBS), contextual word embedding (ELMo) trained on clinical notes and a deep recurrent neural network (BiLSTM-CRF). The task consists in extracting drug mentions and their related information (e.g. dosage, frequency, duration, route, condition). We manually annotated 320 clinical notes extracted from a French clinical data warehouse, to train and evaluate the model. We compared the performances of our approach to standard approaches: rule-based or machine learning only, and classic word embeddings. We evaluated the models using token level recall, precision and F-measure.
Results: Models including RBS, ELMo and BiLSTM reached the best results: overall F-measure of 89.9%. F-measures per category were 95.3% for the medication name, 64.4% for the drug class mentions, 95.3% for the dosage, 92.2% for the frequency, 78.8% for the duration, and 62.2% for the condition of the intake.
Conclusions: Associating expert rules, deep contextualized embedding (ELMo) and deep neural networks improves medication information extraction. Our results reveal a synergy when associating expert knowledge and latent knowledge.

Keywords: medication information; natural language processing; electronic health records; deep learning; rule based system, recurrent neural network; hybrid system

Introduction

In 2017, medication consumption represented 37.8 billion euros in spending in France and 16% of the health budget [1]. Adverse drug reactions (ADR) are an important public health problem, representing a major cause of morbidity and mortality since they were fatal to 0.15% of patients. One third of admissions caused by ADRs are preventable, associated to a poorly reported drug history or

rare adverse events [2,3].
Furthermore, electronic health records (EHR) contain some rich information about drug history that would be valuable to the care of patients (e.g. to prevent interaction with another medication and to track side effects), for epidemiology, or pharmacovigilance [4]. A major hurdle in the use of EHR comes from the format of the data itself. Is has been reported that up to 80 percent of the relevant clinical information is solely present in the form of unstructured text and it represents a major barrier to the secondary use of this type of information [5-6].
A way to overcome this issue is to use natural language processing (NLP) techniques to extract, normalize and restructure drug-related information from clinical texts [6-7] and increase the information available for research and health care. Three approaches have been described for this task: expert knowledge modeling, machine learning and hybrid methods combining both.
The first approach relies on the modelization of the expert's knowledge using rules (i.e. expert rules) such as MedEx or MedlEE based on lexicons or regular expressions [8-11]. Rule-based approaches allow for specific extractions but usually lack of sensitivity and do not perform well on new datasets. Rule-based approaches also require domain experts to design and build the rules and are particularly time-consuming. In addition, expertise is rare and costly, which constitutes a severe bottleneck for this type of method.
The second approach, using machine learning, has been developped in complement of expert approaches. Several authors have provided machine learning-based solutions to extract medication name, dosage, frequency, duration, mode, reason of intake and to detect ADRs [12-13]. Most of systems included a conditional random field (CRF) or a support vector machine (SVM) for medication related information extraction [14-17]. Lample et al. and Sadikin et al. introduced bidirectional long short term memory (LSTM) and associated with CRF (biLSTM-CRF) to complete tasks of name entity recognition (NER) and medication information extraction.[18-19]. Tao et al. used a semi-supervised model [20].
In 2018, several systems were proposed to tackle a n2c2 shared task on medication extraction in electronic health records.[21] Among the best performing teams, BiLSTMs associated with CRFs or not were popular.[22-25] Some systems also used attention mechanisms in addition to LSTMs [26] or combined LSTMs with convolutional neural networks [25]. Other teams combined classic entities extraction systems such as cTakes with classifiers such as support vector machines (SVM).[27] Ensemble approaches, combining multiple classifiers were also proposed.[22-24,28]. The dataset used was constituted of 505 discharge summaries extracted from the MIMIC-III database.[29] This dataset contained 16,225 drug mentions in the training set and a total of 50,951 entity annotations again in the training set.
At the conjonction of machine learning and expert rules, hybrid approaches can leverage the frugality of the expert rules in terms of data needs and the flexibility and generalisability of machine learning. For example, Patrick et al. combined a conditional random field (CRF) to identify the named entities, a support vector machine (SVM) to classify relations and a rule-based context engine to identify medication heading. [30] Dai et al. proposed to extract handcrafted features and feed them to a cascading architecture composed of a CRF and two biLSTM-CRF models [23]. Chen et al. used expert rules and a knowledge base to enrich the text and then a biLSTM with attention to perform the task [26].
All of these approaches were designed for text written in English. To the best of our knowledge, there are only a few studies on French corpora. Deleger et al. used a rule-based system [31]. Lerner et al. developed a hybrid system associating expert rules using terminology and bidirectional gated recurrent units with a CRF [32].
In the recent years, the adoption of word embedding methods has led to a significant increase of performances in many NLP tasks [33]. Word embeddings consist in representing the text using a dense vector representation of the vocabulary. Interestingly, word embeddings are computed using large amounts of unannotated data (e.g. Wikipedia). In static word embeddings, a token is

represented by a static numerical vector. Recently, contextual word embedding methods have appeared such as ELMo [34]. Contextual word embeddings provide a varying representation of the tokens with regards to its context in the text. Contextual word embeddings lead to richer representations and help to improve the performances in clinical concept extraction tasks [35]. Furthermore, associated with the incorporation of semantic information, the results are even better[36].
In this work, we aimed at extracting medication-related information from French clinical narratives in a real world setting (i.e. with documents directly extracted from a clinical data warehouse). Our contribution is two fold: 1) We developed a gold standard dataset of annotated clinical documents in French, along with an annotation guide; 2) we developed an hybrid approach combining an association of knowledge base and expert rules, contextualized word embeddings training on clinical text and a deep learning model based on biLSTM-CRF.

Material and Methods

Data

We leveraged the clinical data warehouse of the Assistance Publique – Hôpitaux de Paris (AP-HP) regrouping data collected from 39 hospitals, to build a dataset of 1M documents [37]. These clinical reports (CR) were randomly selected from medical prescriptions, discharge reports, examinations, observation reports and emergency visits of the clinical data warehouse.

Annotated dataset. We created an annotated dataset for training and evaluation. To this aim, we iterativelly developped an annotation guide during the first phase of annotation. This annotation guide is available as supplementary material and online at the URL https://equipe22.github.io/medExtAnnotation. A small portion of the extracted dataset (320 documents) was manually annotated by three medical doctors using the BRAT annotation tool [38]. The annotation were converted to the IOB standard. Tokens that refers to an entitiy are labelled B-entity_type for the first token and then I-entity_type, tokens outside entities mention are labelled O. We splitted the 320 annotated clinical notes in a training set (n=216), a validation set (n=24) and a test set (n=80).

Knowledge base for drug names. We relied on two French national databases. The French national database of drugs (Base de données des médicaments) and a database from the national medical insurance agency (OpenMedic) [39-40]. From these databases, we created a curated and unified dictionnary of drug mentions.

The corpus can be made available on condition that a research project is accepted by the scientific and ethics committee of the AP-HP health data warehouse (https://recherche.aphp.fr/eds/recherche/).

Methods

In a nutshell, after preprocessing, the text is pre-annotated using a set of expert/hand-crafted rules, then the texts are embedded using contextual word embeddings trained on a large corpus of clinical texts. The pre-annotations and the embedded texts are then fed into a biLSTM-CRF to produce the final annotations as shown in Figure 1.

Definition of the task

We aimed at identifying medication related information in clinical documents in French. We were interested in drug names and a set of attributes related to the drug mentions: dosage, frequency, duration, route, condition of administration. A detailed description of the types of entities extracted is provided in Table 1.

Table 1. Description of the task

Type	Description	Examples
Medication name	Descriptions that denote any medication, active molecule, association or protocol	doliprane, paracetamol, augmentin
Medication class	Descriptions that denote any ATC class or common therapy	ß-Lactam, antibiotherapy
Dosage	Dose or concentration of medication in prescription	3 mg, 2 tablets
Frequency	Frequency of medication administration	3 per day, every morning
Duration	Time range for the administration	3 weeks, until the surgery
Route	Medication administration mode	intravenous, per os
Condition	The event which provokes the administration	if pain, if infection

Preprocessing

We preprocessed the input texts as described in textbox 1.

Textbox 1. Preprocessing of the texts.

- Removing acronym points and replacing decimal points by comma
- Removing break lines added during documents conversion to text
- Removing accents
- Replacing apostrophes by spaces
- Detecting sentence boundaries: remaining points or break lines without transitive verbs, preposition or coordinating conjunctions.
- Detecting word boundaries and tokenization: sequence of alphanumeric characters or a repetition of a unique non-alphanumeric characters

Rule-based module (RBS)

The overall approach was organized as follows: we first identified a drug mention or a drug-class mention with the knowledge-base dictionary using exact-matching. The choice of exact matching for this step was driven by the idea of maximizing the precision of the annotations in this pre-annotation step. Then, using the identified mention as an anchor, we extended the search to the attributes of this mention (i.e. frequency,

dosage, duration, mode of administration and condition of administration) in the area surrounding the seed mention. The attributes were detected using a set of hand-crafted rules using regular expressions. Examples of the rules are described in Supplementary table S1. At this stage, the annotated entities were identified by their position and length relatively to the beginning of the document. For the next steps, the annotations were converted to the IOB standard. The output of the rule based system was used for two different purposes: 1) pre-annotating the documents to speed up the annotation process of the gold standard dataset; 2) serve as extra-features to the input of the deep-learning module.

Deep-learning module

We designed an approach leveraging deep neural networks to accomplish the task. We tested three types of word embeddings: skip-gram [41], FastText embeddings [42] and ELMo embeddings [34]; and two architectures of neural network: biLSTM and biLSTM-CRF.

Embeddings. We evaluated the impact of the use of three different word embeddings on the performance of the model. Our baseline was created using a skip-gram embedding trained on the training set only. We also considered a FastText embedding (skip gram model augmented with sub word information) trained on a corpus of 1M documents. Finally, we used an ELMo embeddings, trained on 100k clinical notes, which are contextualized embeddings computed through the internal states of a large bidirectional language model. Embeddings were kept fixed during model training.

Models. We used a deep recurrent neural network (RNN) composed of long short term memory (LSTM) units [43]. Specifically, we used bidirectional LSTMs (BiLSTMs), which are composed of two concatenated layers of LSTMs: one reading the input sequence forward and another one backward allowing the model to take advantage of the context on the left and the right of a token when computing the latent states. The final prediction layer was either a standard dense layer with softmax or a CRF such as in Lample *et al.* [18].

Implementation and optimization of hyper-parameters

We implemented all the models using the python library Keras and Keras-contrib [44] with a tensorflow backend [45]. We trained our models for 50 epochs, using an ADAM optimizer [46] with a learning rate of 0.001 and early stopping with a patience of 8 epochs . We applied a decrease of learning rate on plateau using a factor of 0.1. For models with a final dense layer, we used a categorical cross entropy loss and a softmax activation. For the models with CRF we used a marginal optimization, and a categorical cross-entropy loss. We tuned the following hyper-parameters using a random search with 15 iterations on the parameter space with the library hyperas:

- Batch size: [64, 128],
- LSTM size: [128, 256, 512],
- dropout before and after LSTM and recurrent dropout: [0.0, 0.1, 0.2, 0.3, 0.5, 0.6, 0.7]

See *supplementary table* S2 for a detailed descriptions of the selected hyper-parameters for the final models.

Evaluation

Models

We compared the performances of RBS only, biLSTM only and RBS+biLSTM. For the two former models, we tested the impact of adding Fasttext embeddings (FT) or ELMo embeddings (ELMo).

Metrics

We considered a token extracted as true positive (TP) if it was annotated with the correct category. A false positive (FP) was falsely annotated with respect to the evaluated class and a false negative (FN) was not annotated or annotated with an incorrect class. We computed the precision, recall and F1-measure to evaluate each model (micro-averaging over all entries):

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1\,score = \frac{2 * (Recall * Precision)}{(Recall + Precision)}$$

We also used the Slot error rate (SER) metric. A slot corresponds to a mention of an entity, i.e., a sequence of B and I tokens of the same class. A deletion is a missing slot. An addition is a slot incorrectly added. A substitution or type error is a class replaced by another class. A frontier error is a token added or removed at the end or the start of the slot [47].

$$SER = \frac{insertion + deletion + 0.5 * (type + frontier)}{Nslots}$$

Results

Annotated dataset

The labeled data set contained 320 clinical notes and 19,957 sentences with 173,796 words. Train, validation and test set included respectively 216 , 24 and 80 clinical notes with 13,737 , 1,373 and 4,847 sentences. *Table 2* summarizes the number of tokens and slots for each class in each data set.

Table 2. Number of slots and tokens for each class per data set.

Label	Train		Validation		Test	
	Tokens	Slots	Tokens	Slots	Tokens	Slots
Medication Name	1,385	1,227	146	143	450	398

Medication Class	309	228	38	30	97	76
Dosage	1,366	761	115	62	606	311
Frequency	1,604	600	142	46	468	184
Duration	161	70	26	13	68	37
Route	95	85	8	8	69	55
Condition	192	61	9	3	89	28

Overall comparisons of the models

Table 3 summarizes the results of the different models. Overall, the best models were the hybrid models combining RBS, text embedding with ELMo and BiLSTM (F-measure: 89.86). It had the lowest SER (0.19) with a minimal deletion rate (0.05)

The biLSTM with baseline embedding had the worst results (F-measure: 73.93). Adding FastText and ELMO trained on external datasets increased the F Score by 14.15 and 9.81 points respectively. Combining RBS and biLSTM increased the F-measure by 14.1 points.

The RBS alone had the higher precision (F-measure 94.67) with the lowest insertion error rate (0.03) and frontier error rate (0.04). It had the second lowest type error rate (0.02) but one of the highest deletion error rate (0.23). Adding BiLSTM and ELMo on top of the rule-based system alone increased the F-measure by 10.45 points.

Table 3. Overall medication component information predictions metrics by models.[a]

Model	**F-measure**	**Precision**	**Recall**	**SER**	**I**	**D**	**T**	**F**
RBS	79.41	94.67	72.28	0.29	0.03	0.23	0.02	0.04
BiLSTM	73.93	83.89	67.57	0.45	0.09	0.25	0.07	0.15
BiLSTM + RBS	83.74	88.46	80.24	0.27	0.08	0.13	0.03	0.09
BiLSTM + FT	88.08	89.48	87.17	0.21	0.07	0.08	0.03	0.09
BiLSTM + FT + RBS	88.18	91.73	85.54	0.21	0.07	0.09	0.01	0.07
BiLSTM + ELMO	88.03	88.81	87.38	0.24	0.1	0.08	0.03	0.1
BiLSTM + ELMO + RBS	89.86	90.83	89.17	0.19	0.09	0.05	0.03	0.08

[a] Models are according to their components: BiLSTM = Bidirectional Long Short Term Memory, ELMo = Embedding for language model, FT = FastText embedding (if nor ELMo or FT is mentioned, then we use a skip gram embedding), RBS = Rule based System

Comparison of the results by type of annotation

Table 4 summarizes the metrics of the different models by type of entities.

The rule-based system alone had the lowest F-measure for every class due to a very low recall (Medication class: 7.22). But it had the highest precision for all classes except for *Medication Name* and *Duration*. Associating the rule-based system to a biLSTM increased *Medication Name* , *Medication class* , *Dosage* and*Condition* metrics (respectively by + 3.13, 3.12, 2.06 and 6.26 of F-measure) but decreased the F-measure for *Frequency* ,*Duration* and *Route* by -1, -3.38 and -2.66.

Table 4. Medication information predictions metrics results by models.

Label	**RBS**			**BiLSTM + ELMo**			**BiLSTM + ELMo + RBS**		
	F	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**
Medication Name	90.31	96.46	84.89	92.2	93.79	90.67	95.33	95.33	95.33
Medication Class	13.33	87.5	7.22	62.3	66.28	58.76	64.36	61.9	67.01
Dosage	90.43	96.62	84.98	92.17	91.13	93.23	95.29	95.52	95.05
Frequency	86.13	98.89	76.28	92.8	93.3	92.31	92.24	93.04	91.45
Duration	48.89	49.25	48.53	82.17	86.89	77.94	78.79	81.25	76.47
Route	47.92	85.19	33.33	75.52	72.97	78.26	72.86	71.83	73.91
Condition	33.64	100	20.22	55.9	62.5	50.56	62.16	77.97	51.69

[a] Models are described by their architecture components: BiLSTM = Bidirectional Long Short Term Memory, ELMo = Embedding for language models, RBS = Rule-based System

Discussion

Principal findings

Our system achieved state-of-the-art performances for the task: an F-measure of 95.33 for medication names and 95.29 for dosage detection with a dataset representing 10% of the size of similar datasets (n2c2 2018 shared task [21]). Combining expert knowledge (RBS) with a deep learning system has increased the global F-measure, precision, recall and SER. The most significant impact was on medication name, medication class and dosage. While the rule-based system alone achieved the best precision and the worst recall, its association with the deep learning models helped to increase recall for all information except condition and increase precision only for medication name, dosage and condition of the intake. Adding a deep learning system with ELMo on top of the rule-based system increased all F-measures and recall. Adding a CRF layer increased the performances on the most frequent entities. On the other entities, models with a CRF layer did not get better results (results in supplementary table). Those results are consistent with the literature [17].

Technical significance

It is interesting to note that leveraging the synergy between expert knowledge and deep learning allowed us to achieve performances comparable to the state-of-the-art with only 10% of the data. Infusing knowledge into deep neural networks will probably be a key element in the future progress of the field. The use of externally trained embeddings is a first step in this direction given that they allow the incorporation of latent knowledge

from large corpora into the models. The impact of contextualized embeddings proves that a more accurate representation is even more important. We can expect some increase in the performances with the more recent language representation approaches such as BERT[48] or XLNET[49]. However, the cost for fitting this type of models in terms of computation, time and data will be a challenge for other languages than english, with lower ressources. Therefore, it will be valuable to leverage other types of representations such as ontologies to infuse knowledge into neural networks. A possible path could be through specific embedding techniques such as Poincaré embeddings[50].

Our approach is highly versatile. It can be transposed to any language, as long as writting expert rules is feasible. We used regular expressions to this end but any rule based is possible. The approach is also transposable to other use cases of information extraction or even text classification.

Clinical significance

The performances achieved by the system open the way towards a large scale use in real-life setting. We are currently developing an implementation to perform the medication information extraction at the scale of our institution. The versatility of the approach will enable its transposition to other types of clinical entities and information.

Related works

Our model performs higher raw results than the best model of the I2B2 2009 medication challenge [30] on token metrics (improvement in medication name, dosage, frequencies and duration token level F-measure: + 5.03, + 4.49, + 4.54, + 28.89 respectively). However, a direct comparison is difficult given that the datasets are different. We trained and evaluated our models on a different French corpus of clinical notes from than usual English corpus such as I2B2 medication challenge or MIMIC corpora [51]; [29]. Also, because of language differences, the annotation guidelines were not strictly identical.

In our corpus, the vast majority of medication name slots contained only one token, so we could approximate phrase-level F-measure by the token F-measure for medication names, and compare with recent study: Tao et al. in 2018 reported a medication F-measure of 90.7 on the I2B2 corpus and we achieved here a F-measure of 95.3.[20] However, our results on route token F-measure was lower. In French clinical datasets, the mentions of mode of administration is less structured and more variable than in English clinical texts. Therefore, it is logical to see lower results on this field and it is consistent with a previous study from Deleger et al. [31]. Moreover we took into consideration the condition of intake and not the reason of intake which is more specific and we added a tag regarding the class name, so the overall F-measures cannot be compared. Comparing with another French corpus, our system raw results higher than Lerner et al. hybrid system's [32] which obtained a token-level F-measure of 90.4 %. But comparision should be considered with caution. The corpus used by Lerner et al., even in the same language, is different from ours as it was from a different source and contained only 147 documents.

Combining a rule-based system with a deep learning model had two major benefits: the synergy between the rules and the machine learning increased the performances and the pre-annotation of the documents with the rules decreased the annotation time. Even if hybrid systems had already proved to be efficient [15,20,30,32,51], combining expert knowledge (rules) and latent knowledge (neural network), demonstrated a synergistic effect by increasing the performances in all metrics.

Limitations and Perspectives

We have several perspectives to continue this work. First, we did not reproduce our study on a more standard corpus such as the i2b2 challenge. We have to redevelop all the expert rules for this English corpus. Second, ELMo was trained on a set of 100,000 French clinical notes from a single hospital [52]. However, even with these limits, using ELMo in the models proved to be efficient. We can anticipate even better results with an ELMo model trained on a larger and more diverse corpus. Finally, our study focuses on recognizing medication information entities without extracting the relations among them. Tao et al. described a way to model the relations by predicting boundaries of utterances that contain related medication entities [20]. We plan to extend this to all types of our corpus sentence, independently of the number of medications mentions. To this end, we will build a multitask model to predict medication fields and relations.

In this multitask model, we will predict also medication event markers such as start, stop, increase, decrease, switch or unique intake of medication. Moreover we could also predict meta-attribute markers: those markers would inform on the experiencer concerned by medication entry, on the temporality (in the past, present or for the future) and on the certainty of (e.g., factual, suggested, hypothetical, conditional, negated or contraindicated, see our annotation guideline [53]).

Conclusion

Associating expert rules, deep contextualized embedding (ELMo) and deep neural networks improves medication information extraction. This association achieved high performances on a heterogeneous corpus of clinical French reports despite a small dataset.

Conflicts of Interest

none declared

Aknowledgments

The authors thank the AP-HP health data warehouse for supporting this work.

Abbreviations

BiLSTM : Bidirectionnal long short term memory

CRF : conditionnal random field

ELMo : Embedding for Language Model

F : F measure

FN : False Negative

FP : False Positive

FT : FastText

LSTM : long short term memory

P : Precision

R : Recall

RBS : Rule-based system

RNN : recurrent neural network

SER : Slot error rate

TP : True Positive

RCT: randomized controlled trial

References

1. Drees. Les depenses de sante en 2017.. 2018.

2. Olivier P, Boulb??s O, Tubery M, Lauque D, Montastruc J-L, Lapeyre-Mestre M. Assessing the Feasibility of Using an Adverse Drug Reaction Preventability Scale in Clinical Practice. Drug Safety [Internet] Springer Science and Business Media LLC; 2002;25(14):1035–1044. [doi: 10.2165/00002018-200225140-00005]

3. Pirmohamed M, James S, Meakin S, Green C, Scott AK, Walley TJ, Farrar K, Park BK, Breckenridge AM. Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients.. BMJ 2004;329:15–9.

4. Zhou L, Mahoney LM, Shakurova A, Goss F, Chang FY, Bates DW, Rocha RA. How many medication orders are entered through free-text in EHRs?–a study on hypoglycemic agents.. AMIA Annu Symp Proc 2012;2012:1079–88.

5. Escudié JB, Jannot AS, Zapletal E, Cohen S, Malamut G, Burgun A, Rance B. Reviewing 741 patients records in two hours with FASTVISU.. AMIA Annu Symp Proc 2015;2015:553–9.

6. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, Forshee R, Walderhaug M, Botsis T. Natural language processing systems for capturing and standardizing unstructured clinical information: A

systematic review.. J Biomed Inform 2017;73:14–29.

7. Gonzalez-Hernandez G, Sarker A, O'Connor K, Savova G. Capturing the Patient's Perspective: a Review of Advances in Natural Language Processing of Health-Related Text. Yearbook of Medical Informatics [Internet] Georg Thieme Verlag KG; 2017;26(01):214–227. [doi: 10.15265/iy-2017-029]

8. Sirohi E, Peissig P. Study of effect of drug lexicons on medication extraction from electronic medical records.. Pac Symp Biocomput 2005;:308–18.

9. Jagannathan V, Mullett CJ, Arbogast JG, Halbritter KA, Yellapragada D, Regulapati S, Bandaru P. Assessment of commercial NLP engines for medication information extraction from dictated clinical notes.. Int J Med Inform 2009;78:284–91.

10. Hyun S, Johnson SB, Bakken S. Exploring the ability of natural language processing to extract data from nursing narratives.. Comput Inform Nurs 2009;27:215–23; quiz 224–5.

11. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives.. J Am Med Inform Assoc 2010;17:19–24.

12. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, Liu S, Zeng Y, Mehrabi S, Sohn S, Liu H. Clinical information extraction applications: A literature review. Journal of Biomedical Informatics [Internet] Elsevier BV; 2018 Jan;77:34–49. [doi: 10.1016/j.jbi.2017.11.011]

13. Jagannatha A, Liu F, Liu W, Yu H. Overview of the First Natural Language Processing Challenge for Extracting Medication Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0). Drug Safety [Internet] Springer Science and Business Media LLC; 2019 Jan;42(1):99–111. [doi: 10.1007/s40264-018-0762-z]

14. Doan S, Collier N, Xu H, Pham HD, Tu MP. Recognition of medication information from discharge summaries using ensembles of classifiers.. BMC Med Inform Decis Mak 2012;12:36.

15. Li Q, Spooner SA, Kaiser M, Lingren N, Robbins J, Lingren T, Tang H, Solti I, Ni Y. An end-to-end hybrid algorithm for automated medication discrepancy detection.. BMC Med Inform Decis Mak 2015;15:37.

16. Zhang Y, Xu J, Chen H, Wang J, Wu Y, Prakasam M, Xu H. Chemical named entity recognition in patents by domain knowledge and unsupervised feature learning.. Database (Oxford) 2016;2016.

17. Lafferty J, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001;

18. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural Architectures for Named Entity Recognition. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies [Internet] San Diego, California: Association

for Computational Linguistics; 2016. p. 260–270. [doi: 10.18653/v1/N16-1030]

19. Sadikin M, Fanany MI, Basaruddin T. A New Data Representation Based on Training Data Characteristics to Extract Drug Name Entity in Medical Text.. Comput Intell Neurosci 2016;2016:3483528.

20. Tao C, Filannino M, Uzuner Ö. FABLE: A Semi-Supervised Prescription Information Extraction System.. AMIA Annu Symp Proc 2018;2018:1534–1543.

21. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. Journal of the American Medical Informatics Association [Internet] Oxford University Press (OUP); 2019 Oct;27(1):3–12. [doi: 10.1093/jamia/ocz166]

22. Christopoulou F, Tran TT, Sahu SK, Miwa M, Ananiadou S. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. Journal of the American Medical Informatics Association [Internet] Oxford University Press (OUP); 2019 Aug;27(1):39–46. [doi: 10.1093/jamia/ocz101]

23. Dai H-J, Su C-H, Wu C-S. Adverse drug event and medication extraction in electronic health records via a cascading architecture with different sequence labeling models and word embeddings. Journal of the American Medical Informatics Association [Internet] Oxford University Press (OUP); 2019 Jul; [doi: 10.1093/jamia/ocz120]

24. Kim Y, Meystre SM. Ensemble methodbased extraction of medication and related information from clinical texts. Journal of the American Medical Informatics Association [Internet] Oxford University Press (OUP); 2019 Jul;27(1):31–38. [doi: 10.1093/jamia/ocz100]

25. Yang X, Bian J, Fang R, Bjarnadottir RI, Hogan WR, Wu Y. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. Journal of the American Medical Informatics Association [Internet] Oxford University Press (OUP); 2019 Aug;27(1):65–72. [doi: 10.1093/jamia/ocz144]

26. Chen L, Gu Y, Ji X, Sun Z, Li H, Gao Y, Huang Y. Extracting medications and associated adverse drug events using a natural language processing system combining knowledge base and deep learning. Journal of the American Medical Informatics Association [Internet] Oxford University Press (OUP); 2019 Oct;27(1):56–64. [doi: 10.1093/jamia/ocz141]

27. Miller T, Geva A, Dligach D. Extracting Adverse Drug Event Information with Minimal Engineering. Proceedings of the 2nd Clinical Natural Language Processing Workshop [Internet] Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019. p. 22–27. [doi: 10.18653/v1/W19-1903]

28. Xu J, Lee H-J, Ji Z, Wang J, Wei Q, Xu H. UTH_CCB System for Adverse Drug Reaction Extraction from Drug Labels at TAC-ADR 2017.. TAC 2017.

29. Johnson AEW, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Scientific data Nature Publishing Group; 2016;3:160035.

30. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge.. J Am Med Inform Assoc 2010;17:524–7.

31. Deléger L, Grouin C, Zweigenbaum P. Extracting medication information from French clinical texts.. Stud Health Technol Inform 2010;160:949–53.

32. Lerner I, Paris N, Tannier X. Terminologies Augmented Recurrent Neural Network Model for Clinical Named Entity Recognition. Journal of Biomedical Informatics [Internet] 2019 Dec;:103356. [doi: 10.1016/j.jbi.2019.103356]

33. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. 2013.

34. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep Contextualized Word Representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) [Internet] Association for Computational Linguistics; 2018; [doi: 10.18653/v1/n18-1202]

35. Zhu H, Paschalidis IC, Tahmasebi A. Clinical Concept Extraction with Contextual Word Embedding. arXiv:181010566 [cs] [Internet] 2018 Nov; Available at: http://arxiv.org/abs/1810.10566

36. Jiang M, Sanger T, Liu X. Combining Contextualized Embeddings and Prior Knowledge for Clinical Named Entity Recognition: Evaluation Study. JMIR Medical Informatics [Internet] 2019;7(4):e14850. [doi: 10.2196/14850]

37. Daniel C, Serre P, Orlova N, Bréant S, Paris N, Griffon N. Initializing a Hospital-Wide Data Quality Program. The AP-HP Experience.. Computer Methods and Programs in Biomedicine [Internet] 2019 Nov;181:104804. [doi: 10.1016/j.cmpb.2018.10.016]

38. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. BRAT: A Web-Based Tool for NLP-Assisted Text Annotation. Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics [Internet] Stroudsburg, PA, USA: Association for Computational Linguistics; 2012. p. 102–107. Available at: http://dl.acm.org/citation.cfm?id=2380921.2380942

39. Accueil - Base de données publique des médicaments [Internet]. http://base-donnees-publique.medicaments.gouv.fr; Available at: http://base-donnees-publique.medicaments.gouv.fr

40. Médicaments remboursés par l'Assurance Maladie - data.gouv.fr [Internet].

https://Data.gouv.fr/fr/datasets/medicaments-rembourses-par-lassurance-maladie; Available at: https://Data.gouv.fr/fr/datasets/medicaments-rembourses-par-lassurance-maladie

41. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and Their Compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. Advances in Neural Information Processing Systems 26 [Internet] Curran Associates, Inc.; 2013. p. 3111–3119. Available at: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

42. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. arXiv preprint arXiv:160704606 [Internet] 2016; Available at: https://arxiv.org/abs/1607.04606

43. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation [Internet] 1997 Nov;9(8):1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

44. Chollet F, others. Keras: The python deep learning library. Astrophysics Source Code Library 2018;

45. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, others. Tensorflow: A system for large-scale machine learning. 12th ⌈USENIX⌋ Symposium on Operating Systems Design and Implementation (⌈OSDI⌋ 16) 2016. p. 265–283.

46. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980 2014;

47. Makhoul J, Kubala F, Schwartz R, Weischedel R, others. Performance measures for information extraction. Proceedings of DARPA broadcast news workshop Herndon, VA; 1999. p. 249–252.

48. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018.

49. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: Generalized Autoregressive Pretraining for Language Understanding. 2019.

50. Agarwal K, Eftimov T, Addanki R, Choudhury S, Tamang S, Rallo R. Snomed2Vec: Random Walk and Poincaré Embeddings of a Clinical Knowledge Base for Healthcare Analytics. 2019.

51. Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. Journal of the American Medical Informatics Association [Internet] Oxford University Press (OUP); 2010 Sep;17(5):514–518. [doi: 10.1136/jamia.2010.003947]

52. Neuraz A, Llanos LC, Burgun A, Rosset S. Natural language understanding for task oriented dialog in the biomedical domain in a low resources context. arXiv preprint arXiv:181109417 2018;

53. Medication extraction annotation guide for french clinical texts [Internet]. https://equipe22.github.io/medExtAnnotation/; Available at: https://equipe22.github.io/medExtAnnotation/

Chapitre 4

PyMedExt : une trousse à outils pour l'annotation de texte et l'échange de données dans le domaine clinique

4.1 Introduction

Comme nous avons pu le voir dans les chapitres précédent, la quantité d'information présente uniquement sous forme de texte libre rédigés par les cliniciens pendant la prise en charge des patients est loin d'être négligeable. Nous avons également pu voir que l'extraction de cette information en routine dans un contexte clinique pourrait aider non seulement les efforts de recherche mais également être utilisé pour guider la décision clinique et la gestion du patient. A travers les années, la communauté du traitement automatique du langage a développé de multiples standards pour échanger et stocker les textes et leurs annotations. Pour n'en citer que quelques uns, CoNLL [172], BioNLP (format standoff) [176], tous deux associés à des tâches partagées de compétitions internationales, BioC, un standard supporté par l'agence américaine National Library of Medicine du National Health Institute [177], ou Universal Dependencies, spécialisé dans la gestion des features linguistiques. Dans le même temps, un grand nombre d'annotateurs et de logiciels de traitement du texte ont été proposé avec des applications allant de l'extraction d'entité nommées cliniques à la détection de la négation, à des classifieurs cliniques, etc. Malgré les efforts de la communauté, aucun standard n'a encore su émerger en terme de format d'échange ou de plateforme d'annotation.

Récemment, la communauté d'informatique médicale a développé des formats d'échange standards, comme Fast Healthcare Interoperability Resources (FHIR) de l'organisation Health Level Seven International (HL7) [178], simplifiant la communication entre les systèmes d'information clinique. De la même façon, des modèles de données communs ont été largement adopté pour stocker les informations cliniques et les résultats des analyses de traitement du texte, le plus connu étant le modèle de données commun de l'Observational Medical Outcomes Partnership [1] (OMOP CDM) [179]. En anglais, des outils avancés de traitement du texte existent, mais ils sont souvent complexe à personaliser et sont en général dépendants d'un unique langage de

1. https://ohdsi.github.io/TheBookOfOhdsi/CommonDataModel.html#fn20

programmation (JAVA si l'on s'intéresse au modèle d'annotation Unstructured Information Management applications (UIMA)) [180] Pour des langues autres que l'anglais, la situation est plus complexe et les ressources sont beaucoup plus limitées.

Nous avons conçu PyMedExt pour faire le pont entre les outils de production des données et les systèmes de traitement de la langue [181]. (Figure 4.1) PyMedExt se décline en trois composant : un format de données, des connecteurs (entre les formats standards) et un outil simple de gestion de workflow. PyMedExt est distribué comme un logiciel qui peut être utilisé pour réaliser des conversions en ligne de commande, ou comme une librairie Python pour un usage plus avancé (e.g., workflow, développement d'annotateurs). PyMedExt est open-source, implémenté en Python 3 et accessible ici : https://github.com/equipe22/pymedext_core

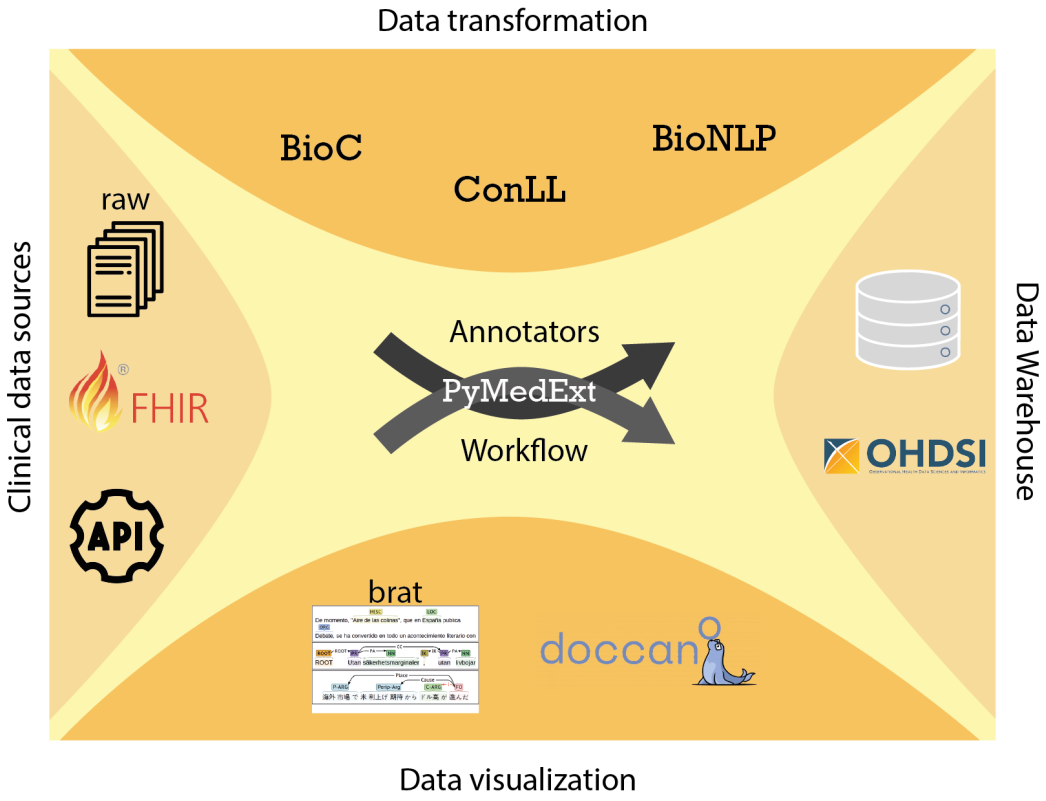


FIGURE 4.1 : Vue générale de PyMedExt

4.2 Les composants de PyMedext

PyMedext comprend trois composants :

1. Le format PyMedExt, une généralisation du format BioC ;
2. Les connecteurs PyMedExt qui permettent de convertir différents formats d'entrée vers la représentation interne et de produire en export une variété de formats standards (BioC, OMOP CDM, BRAT, Doccano) ;
3. le workflow PyMedExt, qui permet la description de workflows de traitement du texte simples.

L'implémentation des différentes classes de PyMedExt sont décrites dans la Figure 4.2

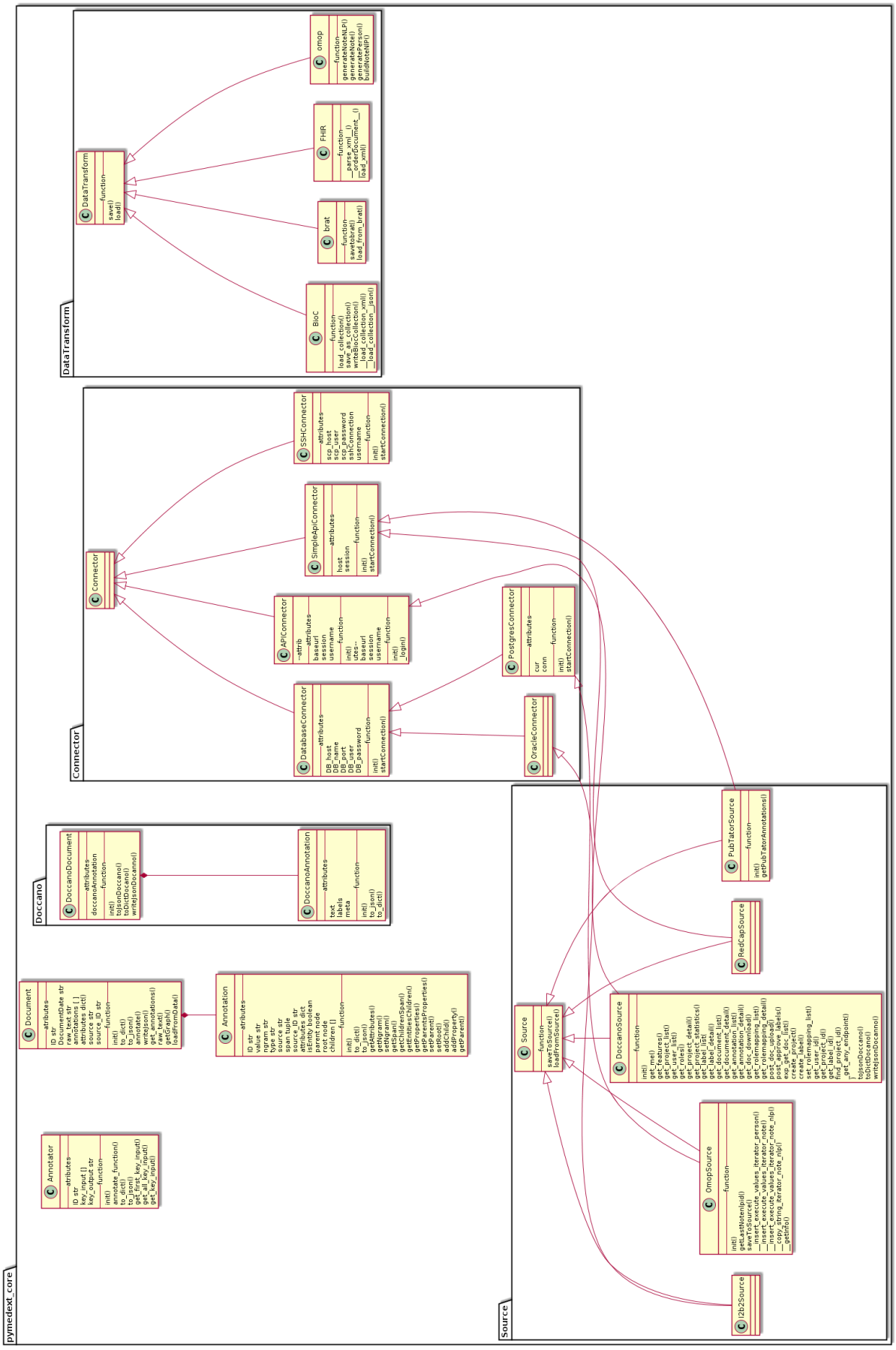


FIGURE 4.2 : Les différentes classes de PyMedExt

4.2.1 Le format PyMedExt

Le coeur du format PyMedExt est un objet de classe `Document` qui encapsule le texte, ses méta-données et ses annotationss. Ce format a été pensé comme une généralisation du format BioC pour permettre une plus grande versatilité.

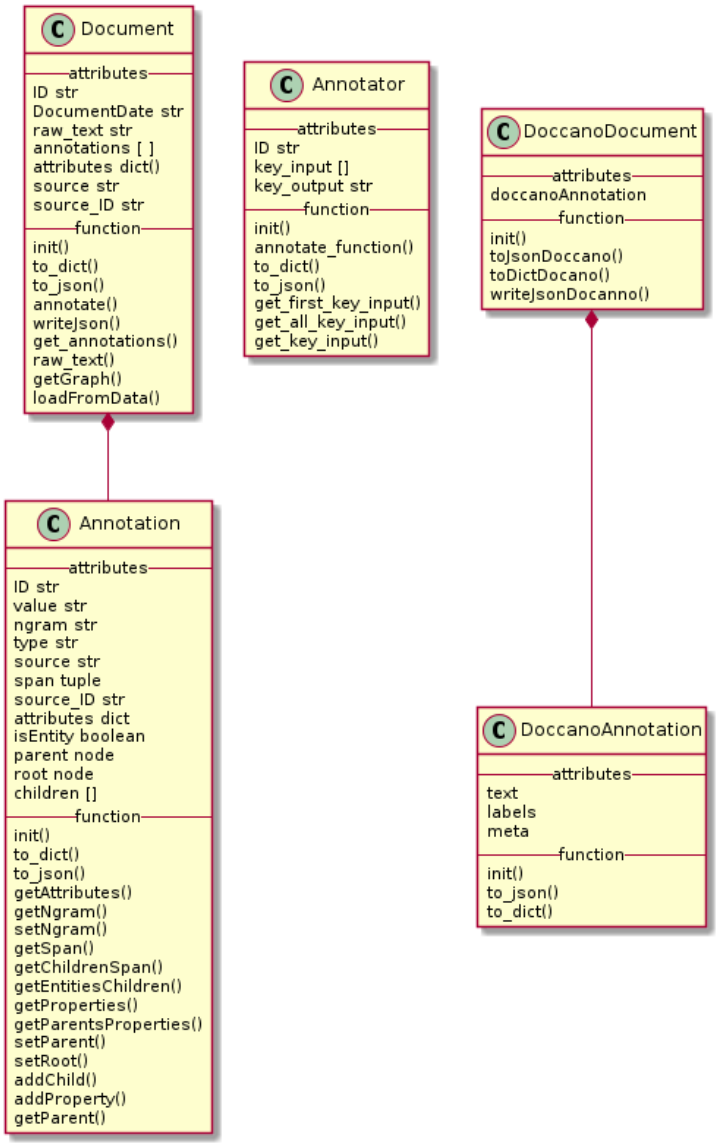


FIGURE 4.3 : La classe Document de PyMedExt

D'abord, nous nous basons sur seulement trois types d'annotations : les entités, les segments et les relations. Chaque élément d'annotation est identifié par un identifiant unique et des informations de provenance y sont associé : identification de l'outil et de la source. Chaque élément peut être associé à des méta-données spécifiques sous forme de dictionnaire clé-valeur. Les entités et les segments sont identifiés en utilisant leur position dans le texte source. La différence entre une entité et un segment tient au fait qu'une entité peut être normalisée (relié à une terminologie standardisée) alors que le segment ne peut pas l'être. Les relations quand à elles sont définies par leur(s) source(s) et leur(s) cible(s) qui peuvent être n à n. Les sources et cibles sont des

entités ou des segments.

Ensuite, nous avons choisi de représenter les entités imbriquées en utilisant des pointeurs grâce à l'identification de la source de chaque annotation. De ce fait, toutes les `Annotations` sont stockées comme des descendants de premier ordre du `Document`. Cette topologie permet la représentation d'annotations avec recouvrement partiel alors qu'une représentation imbriquant les annotations ne le permet pas.

Enfin, nous avons adopté une représentation efficiente des annotations utilisant des arbres d'intervalles pour une localisation rapide et une représentation des annotations sous forme de graph afin de faciliter la propagation des attributs dans les annotations imbriquées et les relations.

L'implémentation du format PyMedExt est basée sur JSON. Nous fournissons également une implémentation Python avec des classes dédiées pour les Documents et les Annotations. Les documents peuvent être organisés en Corpora pour faciliter le traitement en lots et l'organisation des études.

4.2.2 Les connecteurs PyMedExt

PyMedExt vise à simplifier l'utilisation des pipelines NLP dans le cadre clinique. Dans cette optique, nous avons inclus une série de connecteurs pour charger des standards NLP et des standards cliniques vers notre format interne et en sens inverse, pour exporter depuis le format de représentation interne (intégrateurs), vers les standards NLP, des outils d'annotation et de visualisation ou vers des standards cliniques (convertisseurs). Les intégrateurs peuvent prendre en entrée divers formats : NLP (BioC, BioNLP, ConLL-BRAT) ou cliniques (HL7 FHIR). Les convertisseurs transforment les données au format PyMedExt interne vers les différents formats listés ci-dessus et vers le format OMOP CDM (plus précisément les tables Note et Note_NLP). Ces opérations sont transparentes pour l'utilisateur qui pourra choisir d'importer ses données depuis un format, effectuer une série d'opérations d'annotation sur les textes et exporter vers le format de son choix. De plus, pour simplifier l'exploration et la validation d'annotations, le convertisseur pour produire des données compatible avec le logiciel d'annotation BRAT et peu envoyer des documents et annotations vers l'outil Doccano à travers son API.

4.2.3 Le workflow PyMedDext

En plus des mécanismes d'échange de données, PyMedExt vise également à simplifier l'utilisation et la combinaisons d'annotateurs multiples. Nous avons donc inclus une méthode simple pour définir des annotateurs et un framework pour exécuter séquentiellement une série d'annotateurs. Le workflow peut être défini sous la forme d'un graphe dirigé acyclique dans lequel chaque annotateur peut prendre en entrée les sorties d'un ou plusieurs autres annotateurs. En revanche, l'implémentation actuelle ne gère pas automatiquement ces dépendances. De ce fait, la séquence des annotateurs doit être spécifiée ne prenant en compte les dépendances.

Les annotateurs PyMedExt peuvent être développés en utilisant la librarie Python. En pratique, l'utilisateur a juste besoin d'implémenter une classe Annotator et au sein de cette classe, la fonction annotate. Les arguments de cette fonction préciseront la ou les types d'entrées et le type de sortie. Il est également possible d'utiliser des outils déjà implémentés dans d'autres langages ou librairies en utilisant un système d'adaptateur. Des tutoriels sur la création d'annotateurs,

d'adapteurs et de workflows sont disponibles sur github. [2]

4.3 Conclusion

Nous avons décrit ici un outil permettant de faciliter la mise production de pipelines d'extractions d'information en améliorant les importations depuis de sources de données médicales, l'interaction entre différents formats d'annotations et outils de visualisation d'information, ainsi que l'export vers des formats de base de données standardisés. C'est un outil qui manque encore de maturité mais qui, nous allons le voir dans le Chapitre suivant, peut aider à répondre à de vraies questions de santé publique. En effet, dans le prochain chapitre, nous détaillons une étude basée sur le déploiement à grande échelles de pipelines extraction d'information.

2. https://github.com/equipe22/pymedext_core

Chapitre 5

Traitement de la langue naturelle pour une réponse rapide aux maladies émergentes : application au traitement par inhibiteurs calciques pendant la pandémie de COVID-19

La fouille de données biomédicales (biomedical text mining) dans les DPI a souvent été proposé comme méthode pour convertir les données non structurées vers les données structurées nécessaires pour la santé publique. Un des avantages de ces techniques est leur rapidité de développement [182] qui permet de tirer parti des dossiers patients informatisés concernant une maladie nouvelle aussi rapidement qu'ils sont insérés dans le système. Bien que cela ait souvent été suggéré [183], l'occasion ne s'était encore jamais présentée de pouvoir tester cette hypothèse en temps réel. Ainsi, la crise du coronavirus, malgré toutes ses tragédies, présente également l'opportunité d'améliorer l'informatique de santé publique. Ce travail évalue ces possibilités avec une étude de cas concernant l'effet des traitements par inhibiteurs calciques chez les patients hypertendus et leur devenir en cas d'infection par le COVID-19. L'association entre inhibiteurs calciques et l'issue d'une infection par COVID-19 a déjà été suggérée [184], mais n'a encore jamais été explorée dans une grande étude multicentrique.

5.1 Materiels et méthodes

Les données utilisées dans cette étude proviennent des 39 hôpitaux de l'Assistance-Publique Hôpitaux de Paris (APHP). La région parisienne a été très touchée par la première vague de COVID-19 et un des intérêts de l'APHP est qu'elle constitue un réseau d'hôpitaux assez unique, offrant une large diversité de patients. Le 4 mai 2020, l'entrepôt de données spécialisé COVID de l'APHP (EDS-COVID) contenait 84 966 dossiers de patients suspects ou confirmés de COVID-19 (voir la Table 5.3 pour plus de détails). La base EDS-COVID est un entrepôt de données au format du modèle de données commun de OMOP CDM [179], spécialisé pour les patients suspects

ou confirmés COVID-19. Les dossiers informatisés comprennent des données structurées ainsi que des documents en texte libre incluant des notes cliniques et des compte-rendus médicaux. La plupart de ces documents en texte libre ne suivent pas de structure particulière et contiennent différents types d'informations ,*e.g.*, antécédents personnels et familiaux, résultats de laboratoire, historique de traitement, prescriptions médicamenteuses. Cela en fait de très bons candidats pour évaluer les capacités de la fouille de données textuelles. Voici un résumé de la méthode que nous avons utilisé pour traiter les textes clinique (voir la section suivante pour plus de détails) :

- un pré-traitement classique (*i.e.*, nettoyage du texte, détection des phrases) a été appliqué sur l'ensemble du dataset,
- l'extraction des noms de médicaments et des détails de prescription (dose, voie d'administration, fréquence, durée) a été effectuée à l'aire de modèles de deep-learning basés sur des embeddings contextuels de type BERT [34] ($NLP_{medication}$),
- l'extraction de phénotypes spécifiques associés au COVID-19 (*e.g.*, obésité, fumeur), de scores (*e.g.*, IGS2), et de mesures physiologiques (*e.g.*, Body Mass Index), a été effectuée via une liste d'expressions régulières spécialement développées (NLP_{regexp}) ;
- l'extraction de tous signes, symptômes, comorbidités présentes dans le Unified Medical Language System (UMLS) [185], a été effectuée avec l'algorithme quickUMLS [186] (NLP_{UMLS}).

5.1.1 Les étapes du texte clinique à la table OMOP `NOTE_NLP`

Comme décrit dans la Figure 5.1, nous avons appliqué un pré-traitement pour nettoyer les textes avant l'extraction phénotypique.

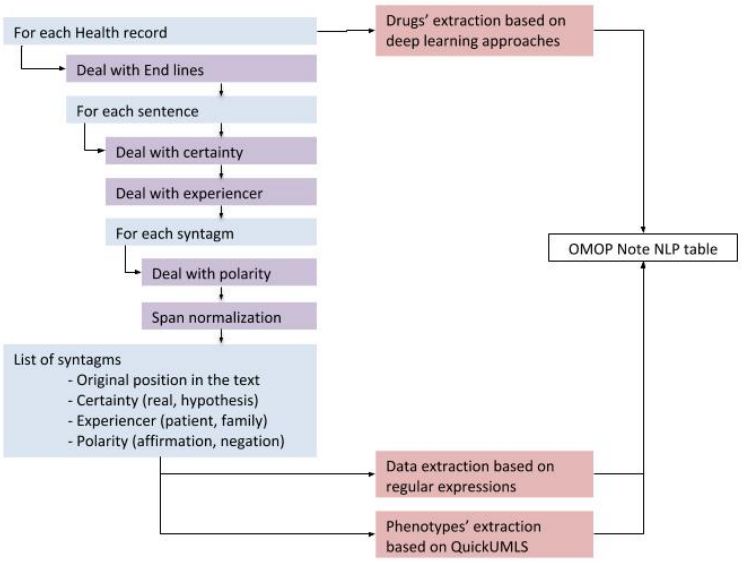


FIGURE 5.1 : Description du pipeline de traitement automatique du language

1. Pour détecter correctement la structure des documents, nous devons gérer les sauts de lignes ajoutés par la conversion des documents pdf en texte.
2. Nous réalisons ensuite le découpage en phrases.

3. Ensuite, les phrases sont classifiées selon la certitude (*i.e.*, `hypothèse` ou `certain`) et le sujet (*i.e.*, histoire `familiale` ou `patient`). Par exemple, la phrase : "en cas de fièvre, prendre 1g de paracétamol." est classée en `hypothèse` et `patient`".
4. Les phrases sont ensuite découpées en syntagmes qui sont classés selon leur polarité (`negation` ou `affirmation`). Les syntagmes sont stockés avec leur position dans le texte.
5. NLP_{regexp} : A partir des syntagmes, les phenotypes sont extraits via des expressions régulières (regexp) définies par des experts. Au total, 60 comorbidités et valeurs ont été définies au format JSON (règle, format, version, critères d'exclusion). Des exemples sont disponibles dans la Figure 5.2. Les regexps ont été développées de manuellement et de manière itérative par des experts en informatique médicale en association avec des médecins. Nous avons évalué leur précision au niveau de la phrase sur un échantillon aléatoire de 100 occurrences par regexp.
6. NLP_{UMLS} : Reprenant les syntagmes de l'étape 4, nous avons également appliqué l'algorithme de QuickUMLS. Cet algorithme permet de détecter des concepts de l'UMLS en utilisant le matching approximatif. Nous avons limité les concepts à extraire au groupe sémantique `Disorders`de l'UMLS version 20199AA.

Ainsi pour chaque méthode, nous obtenons une liste de phénotypes et de scores ainsi que les modifieurs hérités des phrases et syntagmes (*i.e.*, certitude, sujet, polarité) comme décrit dans un travail précédent [187].

7. $NLP_{medication}$: Pour l'extraction des médicaments et de leurs attributs de prescription (*i.e.*, dose, fréquence, durée, condition de prise), nous avons utilisé le modèle d'extraction séquentiel (BERT-BiLSTM-CRF) décrit dans la section 3.2.2 du `Chapitre/@ref`(corpus). Les noms des médicaments sont ensuite normalisés vers la terminologie Anatomical Therapeutic Chemical (ATC) en utilisant un matching approximatif avec le dictionnaire Romedi.[188]
8. L'étape finale a consisté à formater et standardiser toutes les informations extraites pour correspondre à la table `NOTE_NLP` du OMOP CDM. Cela a permis de réintégrer les données directement dans la base EDS-COVID quotidiennement.

```
{"libelle":"BPCO, maladie pulmonaire, pneumopathie, HTAP",
    "regexp":"[^a-z]BPCO[^a-z]|[^a-z]HTAP[^a-
z]|bronchopneumopathie|maladie.{0,5}pulmonaire|pneumopathie|a[sth][sth][sth]me|bronch?o
spasme|a[sth][sth][sth]matique|insuf.{1,2}isance.{0,5}respiratoire.{0,5}chronique|emph.
s[\u00e8e]me|emf.s[\u00e8e]me",
    "id_regexp":"id_regexp_bpco",    "list_cui":"C0024117,C0024115,C0032285,C0020542,C0
004096,C0006266,C0264492,C0034067",
    "regexp_exclude":"",
    "version":"v2",
    "filter_document":"",
    "date_modification":"12\/05\/2020",
    "deprecated":"false",
    "comment":"",
    "refresh":"false"}

{"libelle":"Cancer",
    "regexp":"(?<!pre)(?<!pre )cancer[^a-
z]|tumeur(?!.{0,7}benigne)|carcinome|m[e\u00e9]lanome|n[e\u00e9]oplasie|sarcome",
    "id_regexp":"id_regexp_cancer",
    "list_cui":"C0027651,C1882062,C0006826",
    "regexp_exclude":"",
    "version":"v2",
    "filter_document":"",
    "date_modification":"12\/05\/2020",
    "deprecated":"false",
    "comment":"",
    "refresh":"false"}
```


FIGURE 5.2 : Exemples d'expressions régulières pour l'extraction de phénotypes

Concernant $NLP_{medication}$, le modèle d'extraction d'entités nommées (NER) était un modèle BiLSTM-CRF [124] s'appuyant sur une représentation des tokens utilisant les embeddings contextuels de type "Bidirectional Encoder Representations for Transformers" (BERT) [34]. Les représentations de BERT utilisaient le modèle pré-entraîné "BERT base multilingual" distribué par la librairie HuggingFace `Transformers` [189]. Afin d'améliorer les performances de représentation du modèle BERT pré-entraîné, nous avons affiné l'entraînement (fine-tuned) sur un set de 10 millions de textes cliniques issus de l'entrepôt de données de santé de l'APHP. Pour le modèle de NER, nous avons utilisé l'implémentation de la librairie FLAIR [173] avec deux couches de 1,024 unités pour le LSTM. Nous avons utilisé un optimiseur de type ASGD avec une réduction du pas d'apprentissage sur plateau.

5.1.2 Application clinique : Traitement au long cours par inhibiteurs calciques et devenir des patients lors d'une infection COVID-19.

Le but de ce cas d'usage était d'évaluer les effets potentiels des inhibiteurs calciques (IC) sur la mortalité intra-hospitalière dans le cadre d'une infection par COVID-19.

La raisonnement concernant un effet des inhibiteurs calciques réside dans l'implication du calcium (Ca^{2+}) dans la physiologie des interactions virus-hôte. Le Ca^{2+} est un second messager impliqué dans l'entrée du virus, la réplication des gènes viraux, la maturation du virion et la libération du virus. La modification de l'homéostasie du Ca^{2+} fait partie des stratégies virales pour moduler le comportement des cellules hôtes et générer des dysfonctionnements des organes en leur faveur [190]. La réplication *in vitro* et parfois même *in vivo* de certains virus comme le virus de la grippe A (*influenza virus A*), le virus de la Dengue, le West Nile virus ou encore Ebola a pu

Médicament	Code ATC
Amlodipine	C08CA01
Diltiazem	C05AE03|C08DB01
Felodipine	C08CA02
Isradipine	C08CA03
Lacidipine	C08CA09
Lercanidipine	C08CA13
Manidipine	C08CA11
Nicardipine	C08CA04
Nifedipine	C08CA05
Nitrendipine	C08CA08
Verapamil	C08DA01

TABLE 5.1 : Noms et codes ATC des inhibiteurs calciques

Phénotype	Code CIM10
Cancer	all codes from Chapter II
Diabète	E10, E11, E12, E13, E14
Hypertension artérielle	I10, I15
Obésité	E66

TABLE 5.2 : Noms et codes CIM10 des phénotypes

être inhibée par les IC [191]. Des études préliminaires ont permis de montrer un effet *in vitro* des IC sur la réplication du severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) responsable du COVID-19 [184].

Nous avons comparé les résultats obtenus avec deux sources de données différentes :

1. deux types de données structurées : les codes diagnostics de la classification internationale des maladies version 10 (CIM10) pour les comorbidités et les prescriptions informatisées pour les traitements médicamenteux.
2. informations sur les médicaments et les comorbidités extraites des textes cliniques du dossier patient informatisé via un pipeline NLP.

Le critère d'inclusion dans l'étude était la présence d'une RT-PCR COVID-19 positive. Nous avons considéré qu'un patient était traité par inhibiteur calcique (Table 5.1 s'il y a avait au moins deux mentions (quelle que soit la source de données) dans les six derniers mois. Concernant les comordibités, une seule occurrence était requise pour les codes CIM10 (Table 5.2) et deux occurrences pour le NLP, dans les six derniers mois.

Le critère de jugement principal était la mortalité intra-hospitalière toutes causes. Nous avons utilisé un modèle de Cox multivarié, ajusté sur l'age, le genre, et la présence d'une obésité, d'un diabète et/ou d'un cancer. Le seuil de significativité retenu était 0.05 et tous les tests statistiques étaient bilatéraux. Les analyses ont été réalisées à l'aide du logiciel R statistical software v.3.6.2 [192] avec le package survival [193].

Source	Patients, N=84,966	Documents, N=1,524,057	Data
NLP Médicaments	45,593 (53%)	696,125 (46%)	5,995,945
NLP RegExp	44,498 (52%)	711,900 (46%)	5,449,932
NLP UMLS	44,035 (52%)	833,610 (55%)	19,626,172
Prescriptions structurées	19,791 (23%)	-	826,554
Codes CIM10	38,993 (46%)	-	1,643,819

TABLE 5.3 : Description des données disponibles en fonction des sources

5.2 Resultats

5.2.1 Pipeline TAL

Comme décrit dans la Table 5.3, l'utilisation d'outils TAL a permis d'augmenter, de manière importante, la quantité d'informations, concernant les médicaments et les phénotypes, disponibles pour l'analyse. Le nombre de points de données pour les médicaments a été multiplié par 7.2 ($NLP_{medication}/Structured_{medication}$) et le nombre de phénotypes par 15.2 (($NLP_{RegExp} + NLP_{UMLS})ICD10_{Codes}$). Parmi les 84,966 dossiers présents dans la base EDS-COVID (Table 5.3), 53% des patients avaient des informations sur les médicaments dans les textes cliniques contre seulement 23% dans les champs structurés. Pour les phénotypes spécifiques avec des codes CIM10 existants (Figure 5.3), l'information était disponible uniquement dans le texte libre pour une majorité de patients : 7,133/8,526 (83%) pour le diabète et 2,138/2,871 (74%) pour l'obésité. Certains items étaient absents des données structurées mais ont pu être récupérés via le TAL, par exemple les symptômes COVID-19 spécifiques comme l'agueusie ou l'anosmie, 2,449 et 2,732 patients respectivement.

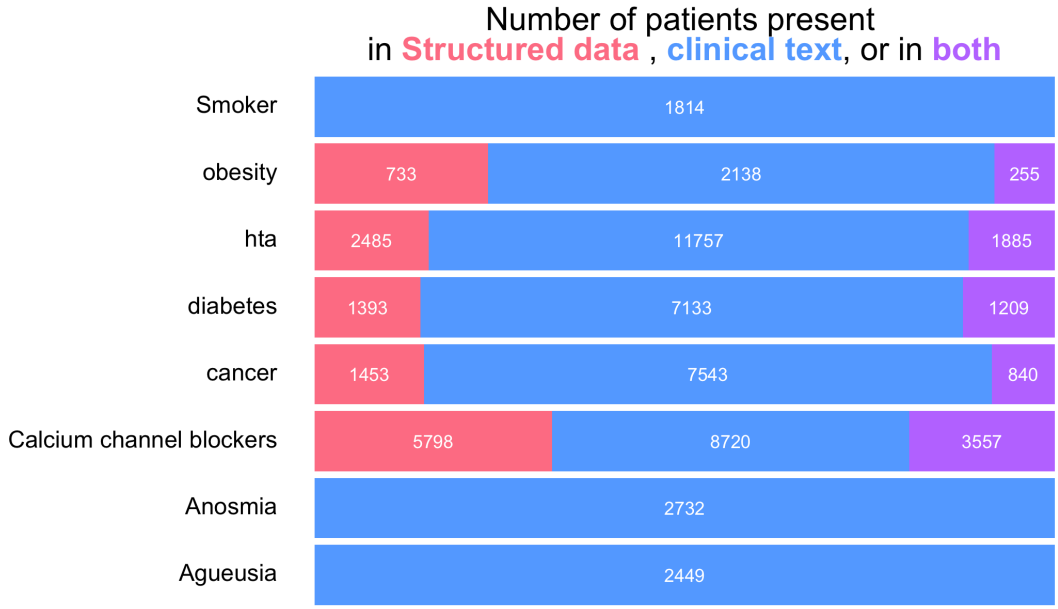


FIGURE 5.3 : Comparaison des données en fonction de leur provenance

En terme de qualité, le modèle $NLP_{medication}$ a montré une F1-mesure brute à 93.8% (91.6%

Entités	Précision (% [90% CI])	Rappel (% [90% CI])	F-mesure (% [90% CI])
Toutes sections			
Nom du médicament	92.1 [89.5-94.5]	95.6 [93.3-97.4]	93.8 [92.1-95.5]
Dose	93.0 [90.0-95.9]	93.9 [91.0-96.5]	93.4 [91.3-95.5]
Fréquence	94.2 [91.4-96.9]	92.6 [88.9-95.6]	93.4 [91.0-95.5]
Sections traitement à l'entrée ou de sortie			
Nom du médicament	94.6 [90.7-98.0]	98.9 [96.8-100.0]	96.7 [94.3-98.7]
Dose	97.6 [94.4-100.0]	97.4 [94.2-100.0]	97.5 [95.2-99.4]
Fréquence	97.3 [93.9-100.0]	97.3 [93.7-100.0]	97.3 [94.9-99.3]

TABLE 5.4 : Performances du modèle d'extraction d'informations sur les médicaments AVANT normalisation des entités.

Entités	Précision (% [90% CI])	Rappel (% [90% CI])	F-mesure (% [90% CI])
Toutes sections			
Nom du médicament	99.2 [98.0-100.0]	85.1 [81.4-88.7]	91.6 [89.3-93.7]
Dose	92.9 [89.9-95.8]	92.4 [89.4-95.2]	92.7 [90.4-94.7]
Fréquence	95.1 [92.1-97.6]	89.6 [85.6-93.2]	92.2 [89.7-94.5]
Sections traitement à l'entrée ou de sortie			
Nom du médicament	100	92.4 [87.3-97.1]	96.0 [93.2-98.5]
Dose	97.4 [94.2-100.0]	97.5 [94.3-100.0]	97.5 [95.1-99.4]
Fréquence	97.2 [93.5-100.0]	93.3 [88.0-98.4]	95.1 [91.7-98.1]

TABLE 5.5 : Performances du modèle d'extraction d'informations sur les médicaments APRES normalisation des entités.

après normalisation) pour l'extraction des noms de médicament sur l'ensemble des sections. En se concentrant sur les sections traitement à l'admission et traitement de sortie, la F1-mesure était 96.7% (96% après normalisation). Les détails des résultats se trouvent dans les Tables 5.4 et 5.5. En ce qui concerne NLP_{regexp}, nous montrons ici les résultats qui ont servi pour le cas d'utilisation : 99% de précision pour l'hypertension, 94% pour l'obésité, 80% pour le diabète et 91% pour le cancer.

5.2.2 Cas d'utilisation

Un total de 3,965 patients ont été inclus en utilisant le TAL contre seulement 1,343 en se restreignant aux données structurées. L'utilisation du TAL a donc permis d'augmenter le nombre de patients dans l'étude par 2.95 (Figure 5.4).

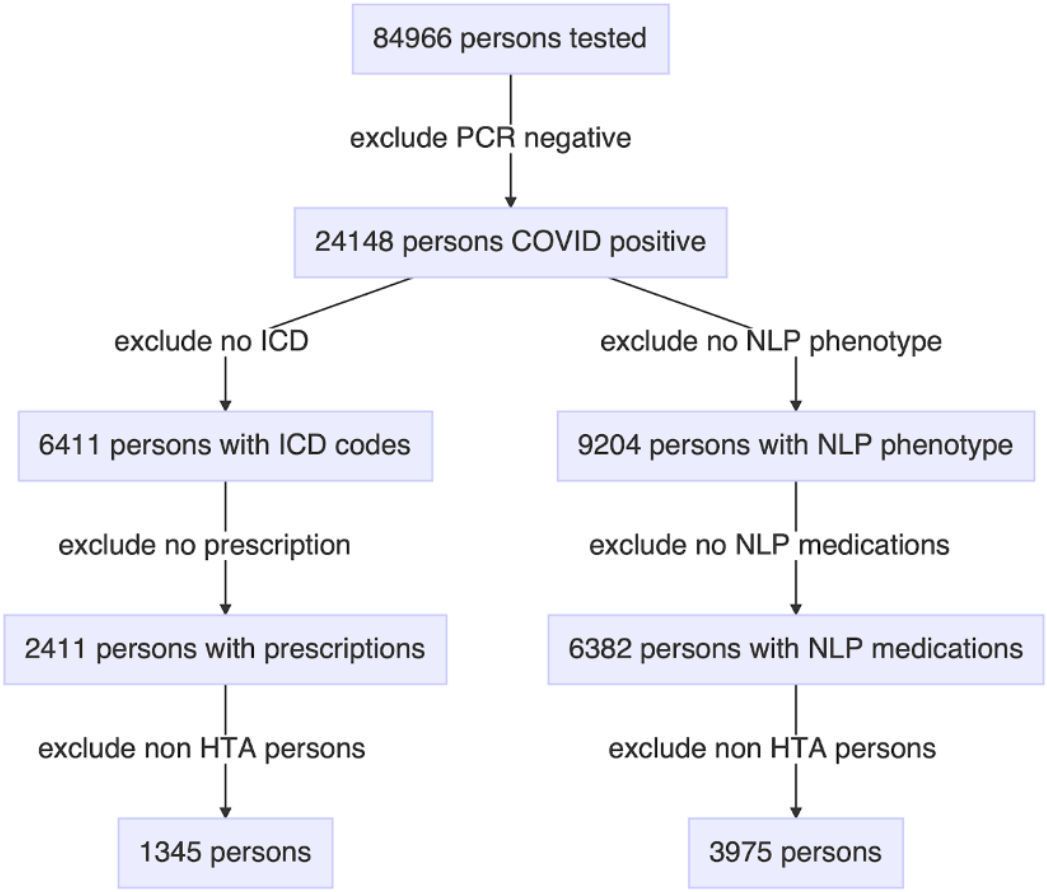


FIGURE 5.4 : Flowchart

Une description détaillée de la population COVID-19 positive avec un historique d'hypertension est disponible dans la Table 5.6. En terme de profondeur temporelle des données sur le traitement par inhibiteurs calciques, la Figure 5.5 montre qu'un plus grand volume d'informations venait du texte plutôt que des données structurées.

Characteristique	TAL, N = 39651	Données structurées, N = 13431
Age		
45-64	1070 (27%)	205 (15%)
18-44	175 (4.4%)	29 (2.2%)
65-74	913 (23%)	252 (19%)
75-84	925 (23%)	392 (29%)
85+	882 (22%)	465 (35%)
Décès	810 (20%)	340 (25%)
Genre		
Femme	1729 (44%)	677 (50%)
Homme	2236 (56%)	666 (50%)
Cancer	886 (22%)	444 (33%)
Diabète	1676 (42%)	560 (42%)
Obésité	518 (13%)	286 (21%)
Inhibiteurs calciques	1846 (47%)	525 (39%)

TABLE 5.6 : Description de la population de l'étude.

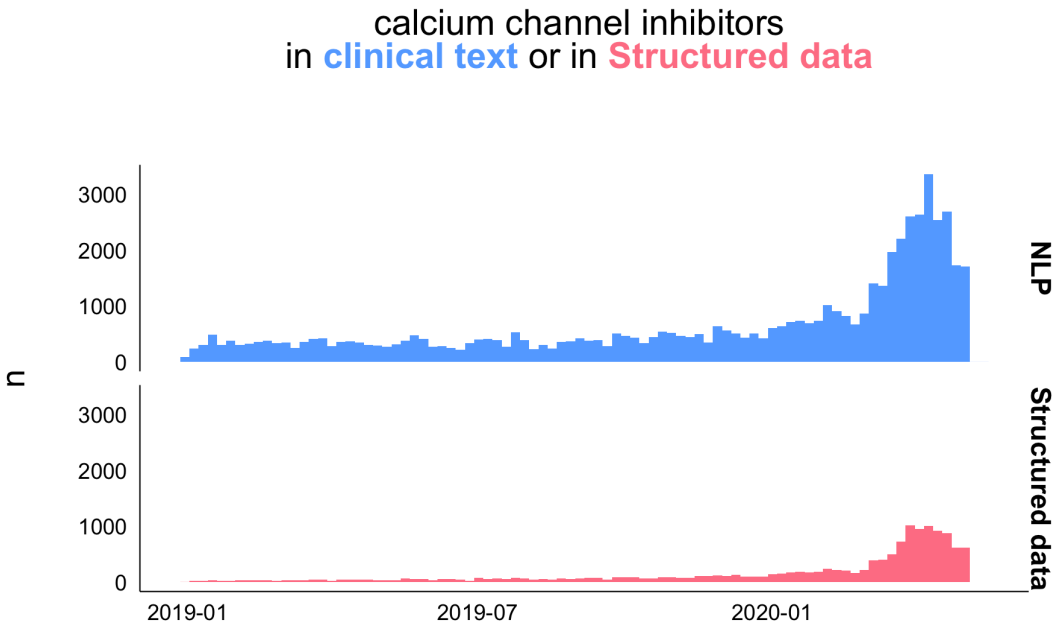


FIGURE 5.5 : Temporalité des données en fonction de leur provenance

Si l'on s'intéresse à présent à l'effet d'un traitement par inhibiteur calcique chez les patients atteints de COVID-19, nous pouvons observer que le hazard ratio ajusté (aHR) est de 0.83 (IC95% 0.67-1.05) en utilisant uniquement les données structurées. Ce résultat n'est pas statistiquement significatif. En revanche, si l'on ajoute les données issues du texte, le aHR devient 0.82 (IC95% 0.71-0.94) et est maintenant statistiquement significatif. La Table 5.7 présente les résultats

	Données structurées			TAL		
Variable	HR[a]	95% CI[b]	p-value	HR	95% CI	p-value
Inhibiteurs calciques	0.83	0.67, 1.05	0.12	0.82	0.71, 0.94	0.005
Age						
45-64	—	—		—	—	
18-44	0.20	0.03, 1.46	0.11	0.35	0.15, 0.80	0.013
65-74	1.50	0.99, 2.27	0.053	1.95	1.54, 2.47	<0.001
75-84	1.68	1.14, 2.48	0.009	2.94	2.35, 3.69	<0.001
85+	2.45	1.66, 3.61	<0.001	3.99	3.16, 5.03	<0.001
Genre						
Femme	—	—		—	—	
Homme	1.59	1.27, 2.00	<0.001	1.53	1.32, 1.77	<0.001
Obésité	1.07	0.81, 1.42	0.6	1.13	0.90, 1.41	0.3
Diabète	1.22	0.98, 1.52	0.080	1.25	1.09, 1.45	0.002
Cancer	1.20	0.96, 1.49	0.11	1.34	1.15, 1.56	<0.001

[a] Hazard Ratio (adjusté) ; [b] Intervale de confiance

TABLE 5.7 : Modèle de Cox multivarié (CI : Interval de confiance, HR hazard ratio).

complets. Des résultats similaires peuvent être observés sur les comorbidités diabète et cancer.

5.3 Discussion

Ce travail cherche à investiguer le potentiel du traitement automatique du langage biomedical dans le contexte de maladies émergentes. Pour ce faire, nous tentons de répondre à la question suivante : Est-ce que le traitement de données textuelles non structurées par des méthodes de traitement automatique du langage permet de produire des informations utiles cliniquement ? Pour répondre à cette question, nous avons utiliser le traitement automatique du langage pour extraire des informations sur l'hypertension et de ses traitements médicamenteux à partir des dossiers informatisés de patients atteints de COVID-19. Les résultats montrent que (1) les pipelines de traitement automatiques du langage peuvent être adaptés rapidement au domaine d'une nouvelle maladie, (2) la qualité des données extraite est suffisante pour fournir des informations utiles, et (3) quand cette information est utilisée pour enrichir les données structurées déjà disponibles, la population d'étude peut être suffisamment augmentée pour faire apparaître des effets de traitement non détectables jusqu'alors.

Plusieurs agences, comme l'agence européenne du médicament, ont mis en évidence les bénéfices de l'utilisation de données de vie réelle pour la recherche, et en particulier pour la génération de preuves supplémentaires et la d'hypothèses [194]. Pendant le pic de la pandémie de COVID-19, le temps disponible pour les cliniciens pour entrer des données dans le système était très réduit. L'informatique médicale est devenue vitale pour gérer la crise dans les hôpitaux et acquérir de nouvelles connaissances sur la maladie. Le pipeline de traitement automatique du langage a été

implémenté en deux semaines, au début de l'épidémie de COVID-19 en France, en s'appuyant sur des travaux précédemment menés à l'AP-HP dans le domaine de l'intelligence artificielle et de la fouille de données textuelles. Plus précisément, la combinaison de développements pré-existants mais non spécifiques (e.g. négation, histoire familiale, hypothèse) avec des extractions sur mesure (i.e. les expressions régulières), nous a permis d'obtenir rapidement des résultats d'une qualité suffisante.

Une soixantaine de projets de recherche exploitant la base EDS-COVID ont été déposés pour évaluation du comité d'éthique local dans les huit premières semaine de l'épidémie. Plus de la moitié nécessitaient des variables comme des symptômes (e.g., agueusie), des signes radiologiques (e.g., crazy pavings), des comorbidités (e.g. obésité) ou l'histoire médicamenteuse (e.g. hydroxychloroquine) qui reposaient sur l'extraction d'information dans les textes cliniques du dossier patient informatisé. [197]

Le cas d'utilisation décrit dans ce travail montre l'impact potentiel de l'utilisation d'informations extraites des textes cliniques des dossiers informatisés pour la recherche sur le COVID-19. Plus précisément, l'étude précédemment décrite aurait eu des conclusions différentes si l'information provenant des textes non structurés avait été exclue. Dans notre exemple, l'ajout des informations issues du TAL n'a pas changé drastiquement la valeur des hazard ratios mais a permis de réduire la taille des intervalles de confiance et d'augmenter la puissance statistique. Il est à noter que cette augmentation de puissance statistique provient essentiellement de l'augmentation du nombre de patients inclus dans l'étude et de la quantité de données disponibles. D'autres analyses seront nécessaires pour évaluer la validité des associations détectées, étant donné que des facteurs confondants peuvent persister et provoquer des résultats faussement positifs. Reproduire l'analyse sur une cohorte de validation et/ou réaliser des tests de falsification [198] pourraient aider à améliorer la validité de ces résultats.

Le travail décrit ici a donné lieu à la publication d'un article original dans la revue Journal of Medical Internet Research (JMIR) :

Neuraz, Antoine, Ivan Lerner, William Digan, Nicolas Paris, Rosy Tsopra, Alice Rogier, David Baudoin, et al. 2020.
"Natural Language Processing for Rapid Response to Emergent Diseases : Case Study of Calcium Channel Blockers and Hypertension in the COVID-19 Pandemic."
Journal of Medical Internet Research 22 (8) : e20773.
<https://doi.org/10.2196/20773>.

JOURNAL OF MEDICAL INTERNET RESEARCH Neuraz et al

Original Paper

Natural Language Processing for Rapid Response to Emergent Diseases: Case Study of Calcium Channel Blockers and Hypertension in the COVID-19 Pandemic

Antoine Neuraz[1,2,3], MD; Ivan Lerner[1,2], MD; William Digan[2,4], MSc; Nicolas Paris[5], MSc; Rosy Tsopra[2,4], MD, PhD; Alice Rogier[2,4], MSc; David Baudoin[2,4], MSc; Kevin Bretonnel Cohen[6], PhD; Anita Burgun[1,2,4], MD, PhD; Nicolas Garcelon[2,7], PhD; Bastien Rance[2,4], PhD; AP-HP/Universities/INSERM COVID-19 Research Collaboration; AP-HP COVID CDR Initiative[8]

[1]Department of Biomedical Informatics, Necker-Enfant Malades Hospital, Assistance Publique – Hôpitaux de Paris (AP-HP), Paris, France
[2]Centre de Recherche des Cordeliers, INSERM UMRS 1138 Team 22, Université de Paris, Paris, France
[3]LIMSI, CNRS, Université Paris Saclay, Orsay, France
[4]Department of Biomedical Informatics, Georges Pompidou European Hospital, Assistance Publique – Hôpitaux de Paris (AP-HP), Paris, France
[5]DSI WIND, Assistance Publique – Hôpitaux de Paris (AP-HP), Paris, France
[6]School of Medicine, University of Colorado, Denver, CO, United States
[7]Institut Imagine, INSERM U1163, Université Paris Descartes, Université de Paris, Paris, France
[8]Please see acknowledgements for list of collaborators

Corresponding Author:
Antoine Neuraz, MD
Department of Biomedical Informatics
Necker-Enfant Malades Hospital
Assistance Publique – Hôpitaux de Paris (AP-HP)
Bat Imagine, Bureau 145
149 rue de Sèvres
Paris, 75015
France
Phone: 33 0624622355
Email: antoine.neuraz@aphp.fr

Abstract

Background: A novel disease poses special challenges for informatics solutions. Biomedical informatics relies for the most part on structured data, which require a preexisting data or knowledge model; however, novel diseases do not have preexisting knowledge models. In an emergent epidemic, language processing can enable rapid conversion of unstructured text to a novel knowledge model. However, although this idea has often been suggested, no opportunity has arisen to actually test it in real time. The current coronavirus disease (COVID-19) pandemic presents such an opportunity.

Objective: The aim of this study was to evaluate the added value of information from clinical text in response to emergent diseases using natural language processing (NLP).

Methods: We explored the effects of long-term treatment by calcium channel blockers on the outcomes of COVID-19 infection in patients with high blood pressure during in-patient hospital stays using two sources of information: data available strictly from structured electronic health records (EHRs) and data available through structured EHRs and text mining.

Results: In this multicenter study involving 39 hospitals, text mining increased the statistical power sufficiently to change a negative result for an adjusted hazard ratio to a positive one. Compared to the baseline structured data, the number of patients available for inclusion in the study increased by 2.95 times, the amount of available information on medications increased by 7.2 times, and the amount of additional phenotypic information increased by 11.9 times.

Conclusions: In our study, use of calcium channel blockers was associated with decreased in-hospital mortality in patients with COVID-19 infection. This finding was obtained by quickly adapting an NLP pipeline to the domain of the novel disease; the adapted pipeline still performed sufficiently to extract useful information. When that information was used to supplement existing

structured data, the sample size could be increased sufficiently to see treatment effects that were not previously statistically detectable.

(J Med Internet Res 2020;22(8):e20773) doi: 10.2196/20773

KEYWORDS

medication information; natural language processing; electronic health records; COVID-19; public health; response; emergent disease; informatics

Introduction

Outbreaks of novel diseases can create enormous strain on public health systems. Since the time of Snow's pioneering work [1] on the epidemiology of the London cholera outbreak of 1854, it has been clear that information is key to the successful abatement of these substantial public health challenges. Currently, health care systems have access to quantities of data that would have been unimaginable in Snow's time. Because these data are in electronic format, they can be manipulated and exploited rapidly. However, a novel disease poses special challenges for informatics solutions. Biomedical informatics relies for the most part on structured data; structured data require a preexisting data or knowledge model; and a novel disease will not have a preexisting knowledge model. This poses a formidable obstacle to leveraging informatics solutions to address the type of public health crisis the world is facing at the time of writing. One solution to the lack of structured information is natural language processing (NLP).

Biomedical text mining, or the use of textual data, in electronic health records (EHRs) has often been proposed as a method for converting unstructured data to the structured data that is needed in public health informatics. One of the advantages of biomedical text mining is that it can be developed rapidly [2], which can permit the leveraging of electronic health records of patients with a novel disease as quickly as they are entered into the EHR. However, although this has often been suggested [3], there has never been an opportunity to actually test that claim in real time. Thus, the current novel coronavirus disease (COVID-19) pandemic, with all of its challenges, presents an opportunity to advance the state of public health informatics. In this paper, we tested this possibility with a case study on the effects of use of calcium channel blockers (CCBs) in patients with high blood pressure on the risk of death from COVID-19 infection. An association between CCB and the outcome of COVID-19 infection has already been suggested [4] but has not previously been explored in a large multicenter clinical study.

Methods

Data Source and NLP Pipeline

The data used in this study were obtained from 39 different hospitals in the Paris metropolitan area in the Assistance Publique – Hôpitaux de Paris (AP-HP) system. Focusing on this region of the country and on a large number of hospitals afforded a diversity of patient demographics that would not be available in most other parts of the country. As of May 4, 2020, the Entrepôt de Données de Santé (EDS)-COVID data set contained 84,966 electronic records of suspected or confirmed patients with COVID-19 (see Table 1 for further details on the data set). The records comprise structured fields and free text documents, including clinical notes and narratives. Most of the textual documents do not follow a specific structure and contain different types of patient information, such as patient history, family history, laboratory results, drug history, and prescriptions. Therefore, they represent an excellent test case for the real abilities of text mining. We used the following pipeline:

- Typical preprocessing steps (ie, text cleaning and sentence detection) were applied to the full data set (see Multimedia Appendix 1 for a detailed description).
- Drug names and details of administration (dose, route of administration, frequency, and duration) were extracted via a deep learning approach based on bidirectional encoder representations from transformers (BERT) contextual embeddings [5] (NLP Medication).
- Specific phenotypes associated with COVID-19 (eg, obesity, smoking status), scores (eg, sequential organ failure assessment score) and physiological measures (eg, BMI), were extracted via a list of 60 regular expressions (NLP RegExp).
- All signs, symptoms, and comorbidities included in the Unified Medical Language System (UMLS) [6] were extracted with the quickUMLS algorithm [7] (NLP UMLS).

A visual depiction of the pipeline is provided in Multimedia Appendix 2.

The NLP medication extraction model was a bidirectional long short-term memory with a conditional random field (BiLSTM-CRF) [8] layer on top of a vector representation of tokens using BERT [5]. We fine-tuned multilingual BERT on a set of 10 million clinical texts from EHRs. The model was trained on the APMed corpus, a manually annotated corpus of French clinical texts described in [9]. We used the FLAIR [10] implementation with 2 layers of 1024 units for the LSTMs with an asynchronous stochastic gradient descent (ASGD) optimizer and a reduction of the learning rate on plateau.

The NLP regular expression for the extraction of specific phenotypes was a set of 60 regular expressions developed manually and iteratively by medical informatics experts and physicians. We evaluated their precision at the sentence level using a random sample of 100 positive sentences for each regular expression. Examples of these expressions can be found in Multimedia Appendix 3.

All the terms extracted by the NLP pipeline, regardless of the method, were automatically annotated according to their modality (negated or hypothetical) and experiencer in the text, as described in previous work [11]. The outputs of the NLP pipeline were normalized to the Observational Medical

Outcomes Partnership (OMOP) common data model (CDM) [12] and were fed back to the database system on a daily basis.

Data Availability

Data supporting this study can be made available on request, on condition that the research project is accepted by the scientific and ethics committee of the AP-HP health data warehouse [13].

Clinical Application: Long-Term CCB Use and Outcomes of COVID-19 in Patients With High Blood Pressure

The clinical goal of this case study was to evaluate the potential effects of CCBs on in-hospital mortality related to COVID-19 [4]. To achieve this goal, we used two different sources of data. The first source was two elements of structured data: International Classification of Disease, Tenth Revision (ICD-10) codes and medication prescriptions from an electronic prescription system. The second source was information on medications and comorbidities extracted by the NLP pipeline from nonstructured fields in the EHR. The inclusion criterion for patients was COVID-19 disease confirmed by reverse transcriptase–polymerase chain reaction (RT-PCR).

We considered a patient as receiving long-term treatment with CCBs (Multimedia Appendix 4) if there were at least two mentions (in structured data or extracted with NLP, respectively) in the last 6 months. We qualified cases as having comorbidities through one occurrence of an ICD-10 code (Multimedia Appendix 5) or two NLP mentions in the last 6 months.

The measured outcome was in-hospital mortality. We used a multivariate Cox proportional hazard model [14] that was adjusted according to age, gender, and the presence of obesity, diabetes, and cancer. The level of significance was set as $P=.05$, and all statistical tests were two-sided. We used R statistical software v.3.6.2 (R Project) with the Survival package.

Results

NLP Pipeline

As Table 1 shows, NLP markedly expanded the quantity of medication and phenotype information available for the analysis. The number of data points for medication increased by 7.2 times (*NLP medication*)/(*structured medication*), and the number of phenotypes increased by 15.2 times ((*NLP RegExp* + *NLP UMLS*)/(*ICD-10 codes*). Among the 84,966 patients with records present in the EDS-COVID cohort (Table 1), 45,593 (53.7%) contained drug information in their narrative EHR documents, whereas only 19,791 (23.3%) of the patients had medication information available in the structured fields in the EHR.

For specific phenotypes with existing ICD-10 codes (Figure 1), information was only available in clinical free-text fields for the majority of patients: 7133/8526 (60.2%) for diabetes, and 2138/2871 (74.5%) for obesity. Some items were absent from the structured data but could be recovered using the NLP extraction pipeline, such as COVID-19–specific symptoms such as ageusia (2449 patients) and anosmia (2732 patients).

In terms of quality, the extraction of medication names showed an F1 score of 93.8% (91.6% after normalization) in all sections. When focusing on the admission and discharge treatment sections, the F1 score was 96.7% (96.0% after normalization). The detailed results are shown in Multimedia Appendix 6. Regarding the phenotypes extracted by regular expressions in our case study, hypertension showed a precision of 99%, and obesity, diabetes, and cancer showed precisions of 94%, 80%, and 91%, respectively.

Table 1. Description of the information extracted using the NLP pipeline in the EDS-COVID cohort (N=84,966).

Source	Patient records (N=84,966), n (%)	Documents (N=1,524,057), n (%)	Data points, n
NLP[a] Medication	45,593 (53.7)	696,125 (45.7)	5,995,945
NLP RegExp[b]	44,498 (52.4)	711,900 (46.7)	5,449,932
NLP UMLS[c]	44,035 (51.8)	833,610 (54.7)	19,626,172
Structured medication	19,791 (23.3)	N/A[d]	826,554
ICD-10[e] codes	38,993 (45.9)	N/A	1,643,819

[a]NLP: natural language processing.

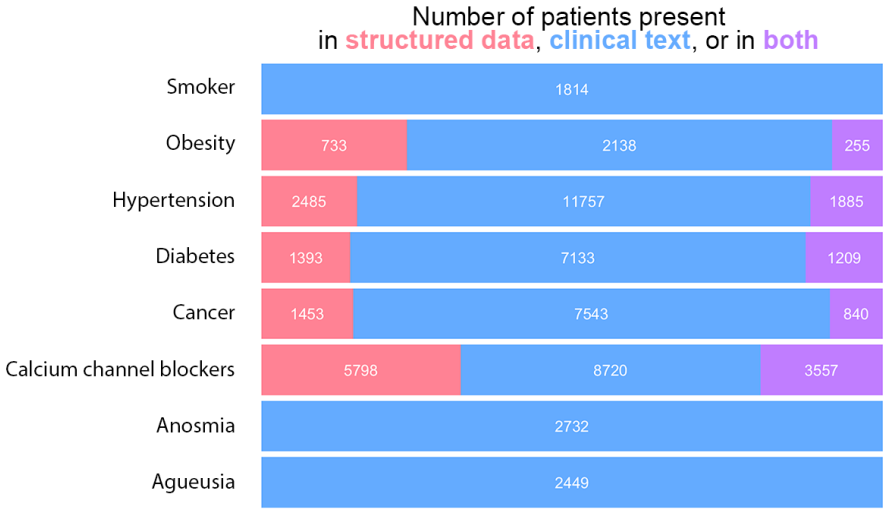
[b]RegExp: regular expression.

[c]UMLS: Unified Medical Language System.

[d]N/A: not applicable.

[e]ICD-10: International Classification of Disease, Tenth Revision.

Figure 1. Quantity of patients with information for a selection of items depending on the source of data.



Case Study

Of the 84,966 total patients, 3965 (4.7%) were included using the NLP pipeline, of which only 1343 (15.9%) could be included if the study were limited to the use of structured data; this increased the number of patients added for the case study increased by 2.95 times (Multimedia Appendix 7). A detailed description of the population of patients who tested positive for COVID-19 with a history of high blood pressure can be found in Multimedia Appendix 8). In terms of the temporal depth of CCB treatment information, Figure 2 shows that a higher volume of information was obtained from text fields compared to structured data.

When using only structured data, we observed an adjusted hazard ratio (aHR) of 0.83 (95% CI 0.67-1.05) for treatment with CCBs; this result was not statistically significant (P=.12). When including NLP data, the aHR became 0.82 (95% CI 0.71-0.94), which represents a statistically significant reduction of the risk of death (P=.005). Similar results can be observed that support an increased risk of mortality with the presence of diabetes and cancer as comorbidities (Table 2).

Figure 2. Quantity of information about calcium channel blockers for the two data sources over time. NLP: natural language processing.

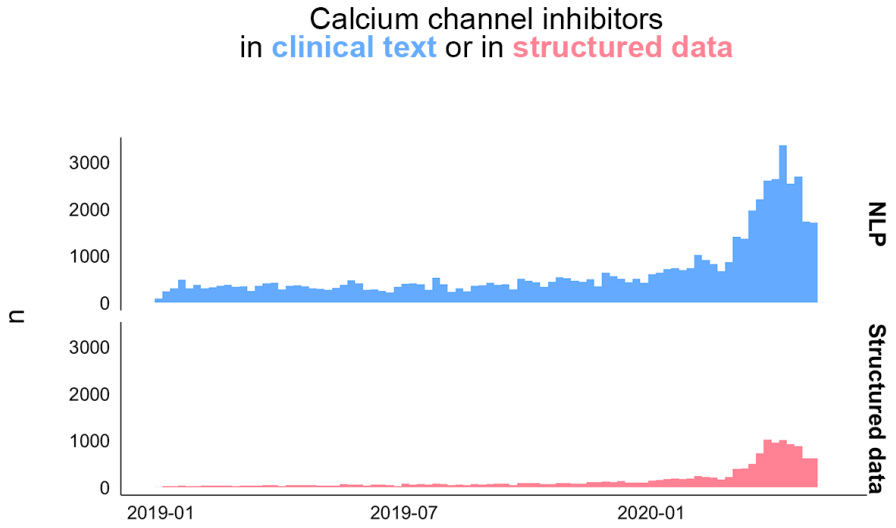


Table 2. Results of the multivariate Cox survival model.

Characteristic	Structured data			NLP[a]		
	aHR[b]	95% CI	*P* value	HR[c]	95% CI	*P* value
Calcium channel blockers	0.83	0.67-1.05	.12	0.82	0.71-0.94	.005
Age (years)						
45-64	Reference	N/A[d]	N/A	N/A	N/A	N/A
18-44	0.20	0.03-1.46	.11	0.35	0.15-0.80	.01
65-74	1.50	0.99-2.27	.053	1.95	1.54-2.47	<.001
75-84	1.68	1.14-2.48	.009	2.94	2.35-3.69	<.001
85+	2.45	1.66-3.61	<.001	3.99	3.16-5.03	<.001
Gender						
Female	Reference	N/A	N/A	N/A	N/A	N/A
Male	1.59	1.27-2.00	<.001	1.53	1.32-1.77	<.001
Obesity	1.07	0.81-1.42	.60	1.13	0.90-1.41	.30
Diabetes	1.22	0.98-1.52	.08	1.25	1.09-1.45	.002
Cancer	1.20	0.96-1.49	.11	1.34	1.15-1.56	<.001

[a]NLP: natural language processing.
[b]aHR: adjusted hazard ratio.
[c]HR: hazard ratio.
[d]N/A: not applicable.

Discussion

In this paper, we investigated the potential utility of biomedical NLP in the context of a rapidly emerging novel disease. To do this, we asked a specific question: Does the leveraging of unstructured textual information via NLP yield clinically actionable information? To answer this question, we used NLP to extract information about hypertension and a medication for treating it from the EHRs of patients with COVID-19. The results showed that an NLP pipeline can be adapted quickly to the domain of a novel disease, it can perform well enough to extract useful information, and when that information is used to supplement the structured data that is already available, the sample size can be increased sufficiently to see treatment effects that were not previously statistically detectable.

Several agencies, notably the European Medicines Agency, have highlighted the benefits of using real-world data for research, in particular for the generation of complementary evidence and new hypotheses [15]. During the peak of the COVID-19 pandemic, the time available for clinicians to enter EHR data was greatly reduced. Medical informatics became vital to manage the crisis in hospitals and acquire better knowledge of the disease. The NLP pipeline was implemented within two weeks at the beginning of the COVID-19 epidemic in France, building on previous developments in artificial intelligence and text mining at AP-HP. More specifically, combining nonspecific preexisting developments (eg, negation, family history, and hypothesis detection) to tailored extractions (ie, regular expressions) allowed us to obtain rapid results of sufficient quality.

Approximately 60 internal research projects exploring EDS-COVID data were submitted for Institutional Review Board approval within the first eight weeks of COVID-19 epidemic. More than half of these projects studied variables such as symptoms (eg, ageusia), radiological signs (eg, crazy paving), comorbidities (eg, obesity), and drug history (eg, hydroxychloroquine), requiring extraction of information from narrative reports in EHRs.

The case study described in this paper shows the possible impact of using information extracted from text in the EHR for COVID-19 research. More precisely, the conclusions of the above study would have been different if information from unstructured fields had been excluded. In our case study, the addition of information from NLP did not dramatically change the hazard ratio from the analyses; however, it allowed us to include more patients and therefore narrowed the CIs and increased the statistical power. Note that the increased statistical power is mainly due to the increase in the number of patients included and the quantity of data available. Further analyses are required to assess the validity of the associations detected here, given that some confounding biases may remain and provoke false positive results. Reproducing the analysis with an external population or performing falsification testing [16] could help improve the validity of these findings.

Acknowledgments

The authors thank the EDS AP-HP COVID consortium integrating the AP-HP Health Data Warehouse team as well as all the AP-HP staff and volunteers who contributed to the implementation of the EDS-COVID database and operating solutions for the database. The authors would like to acknowledge John Bennett for his thorough editing. This work was supported by state funding from the French National Research Agency (Agence Nationale de la Recherche, ANR) under the “Investissements d’Avenir” program (reference: ANR-10-IAHU-01) and an ANR PractikPharma grant (ANR-15-CE23-0028). The collaborators associated with AP-HP/Universities/INSERM COVID-19 Research Collaboration: AP-HP COVID CDR Initiative, Paris, France, are as follows: Pierre-Yves Ancel, Alain Bauchet, Nathanaël Beeker, Vincent Benoit, Mélodie Bernaux, Ali Bellamine, Romain Bey, Aurélie Bourmaud, Stéphane Breant, Anita Burgun, Fabrice Carrat, Charlotte Caucheteux, Julien Champ, Sylvie Cormont, Christel Daniel, Julien Dubiel, Catherine Duclos, Loic Esteve, Marie Frank, Nicolas Garcelon, Alexandre Gramfort, Nicolas Griffon, Olivier Grisel, Martin Guilbaud, Claire Hassen-Khodja, François Hemery, Martin Hilka, Anne Sophie Jannot, Jerome Lambert, Richard Layese, Judith Leblanc, Léo Lebouter, Guillaume Lemaitre, Damien Leprovost, Ivan Lerner, Kankoe Levi Sallah, Aurélien Maire, Marie-France Mamzer, Patricia Martel, Arthur Mensch, Thomas Moreau, Antoine Neuraz, Nina Orlova, Nicolas Paris, Bastien Rance, Hélène Ravera, Antoine Rozes, Elisa Salamanca, Arnaud Sandrin, Patricia Serre, Xavier Tannier, Jean-Marc Treluyer, Damien van Gysel, Gaël Varoquaux, Jill Jen Vie, Maxime Wack, Perceval Wajsburt, Demian Wassermann and Eric Zapletal.

Authors' Contributions

AN, IL, AB, NG, and BR contributed to the conception or design of the work. AN, IL, WD, NP, RT, NG, and BR acquired, analyzed, or interpreted the data. AN, IL, WD, NP, AR, DB, NG, and BR created the new software used in the work. AN, IL, AB, NG, RT, BR, and KBC drafted the work or substantively revised it.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary methods.
[DOCX File , 14 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Description of the natural language processing pipeline.
[DOCX File , 53 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Examples of regular expression for the extraction of phenotypes.
[DOCX File , 13 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Definition of calcium channel blockers (name, ATC number).
[DOCX File , 12 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Definition of phenotypes (name, ICD0-10 code).
[DOCX File , 12 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Performance of the medication information extraction model before and after normalization of the entities.
[DOCX File , 14 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Flowchart of the use case: patients who tested positive for COVID-19 who have hypertension.
[DOCX File , 227 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Characteristics of the population of COVID positive patients with hypertension in EDS-COVID.
[DOCX File , 13 KB-Multimedia Appendix 8]

References

1. Snow J. On the Mode of Communication of Cholera. London, UK: Wilson and Ogilvy; 1855.
2. Chapman W, Dowling J, Ivanov O, Gesteland P, Olszewski R, Espino J, et al. Evaluating natural language processing applications applied to outbreak and disease surveillance. In: Proceedings of 36th symposium on the interface: computing science and statistics 2004. 2004 Presented at: 36th Symposium on the Interface: Computing Science and Statistics 2004; May 26-29, 2004; Baltimore, MD.
3. Elkin PL, Froehling DA, Wahner-Roedler DL, Brown SH, Bailey KR. Comparison of natural language processing biosurveillance methods for identifying influenza from encounter notes. Ann Intern Med 2012 Jan 03;156(1 Pt 1):11-18. [doi: 10.7326/0003-4819-156-1-201201030-00003] [Medline: 22213490]
4. Zhang L, Sun Y, Zeng H, Peng Y, Jiang X, Shang W, et al. Calcium channel blocker amlodipine besylate is associated with reduced case fatality rate of COVID-19 patients with hypertension. medRxiv 2020 Apr 14:preprint. [doi: 10.1101/2020.04.08.20047134]
5. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXivcs. 2018 Oct 10. URL: http://arxiv.org/abs/1810.04805 [accessed 2018-11-17]
6. Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 2018 Feb 06;32(04):281-291. [doi: 10.1055/s-0038-1634945]
7. Okazaki N, Tsujii J. Simple and Efficient Algorithm for Approximate Dictionary Matching. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010).: Coling 2010 Organizing Committee; 2010 Presented at: 23rd International Conference on Computational Linguistics (Coling 2010); August 2010; Beijing, China URL: https://www.aclweb.org/anthology/C10-1096
8. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural Architectures for Named Entity Recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.: Association for Computational Linguistics; 2016 Presented at: 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2016; San Diego, CA p. A. [doi: 10.18653/v1/n16-1030]
9. Jouffroy J, Feldman S, Lerner I, Rance B, Burgun A, Neuraz A. MedExt: combining expert knowledge and deep learning for medication extraction from French clinical texts. ResearchGate 2020 Jan:preprint. [doi: 10.2196/preprints.17934]

10. Akbik A, Bergmann T, Blythe D, Rasul K, Schweter S, Vollgraf R. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations) Internet Minneapolis, Minnesota: Association for Computational Linguistics; 2019 Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations); June 2019; Minneapolis, MI. [doi: 10.18653/v1/n19-4010]
11. Garcelon N, Neuraz A, Benoit V, Salomon R, Burgun A. Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse. J Am Med Inform Assoc 2017 May 01;24(3):607-613. [doi: 10.1093/jamia/ocw144] [Medline: 28339516]
12. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
13. Soumettre un projet de recherche au Comité Scientifique et Ethique de l'Entrepôt de Données de Santé. Assistance Publique — Hôpitaux de Paris. URL: https://recherche.aphp.fr/eds/recherche/ [accessed 2020-08-11]
14. Cox DR. Regression Models and Life-Tables. J R Stat Soc Series B Stat Methodol 2018 Dec 05;34(2):187-202. [doi: 10.1111/j.2517-6161.1972.tb00899.x]
15. EMA Regulatory Science to 2025: Strategic reflection. European Medicines Agency. 2018. URL: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/ema-regulatory-science-2025-strategic-reflection_en.pdf [accessed 2020-08-11]
16. Pizer SD. Falsification Testing of Instrumental Variables Methods for Comparative Effectiveness Research. Health Serv Res 2016 Apr;51(2):790-811 [FREE Full text] [doi: 10.1111/1475-6773.12355] [Medline: 26293167]

Abbreviations

aHR: adjusted hazard ratio
AP-HP: Assistance Publique – Hôpitaux de Paris
ASGD: asynchronous stochastic gradient descent
BiLSTM-CRF: bidirectional long short-term memory with a conditional random field
CCB: calcium channel blocker
CDM: common data model
COVID-19: coronavirus disease
EDS: Entrepôt de Données de Santé
EHR: electronic health record
ICD-10: International Classification of Disease, Tenth Revision
NLP: natural language processing
RT-PCR: reverse transcriptase–polymerase chain reaction
OMOP: Observational Medical Outcomes Partnership
UMLS: Unified Medical Language System

Edited by G Eysenbach; submitted 02.06.20; peer-reviewed by H Kalicoglu, S Zheng, D Pförringer, N Shah; comments to author 23.06.20; revised version received 02.07.20; accepted 26.07.20; published 14.08.20

Please cite as:
Neuraz A, Lerner I, Digan W, Paris N, Tsopra R, Rogier A, Baudoin D, Cohen KB, Burgun A, Garcelon N, Rance B, AP-HP/Universities/INSERM COVID-19 Research Collaboration; AP-HP COVID CDR Initiative
Natural Language Processing for Rapid Response to Emergent Diseases: Case Study of Calcium Channel Blockers and Hypertension in the COVID-19 Pandemic
J Med Internet Res 2020;22(8):e20773
URL: http://www.jmir.org/2020/8/e20773/
doi: 10.2196/20773
PMID:

Chapitre 6

Conclusion

Tout d'abord, dans le cadre des systèmes de compréhension de la langue naturelle pour interroger le DPI, nous avons montré l'intérêt de l'augmentation de données et de l'incorporation de connaissances latentes pour l'entraînement de modèles en l'absence de jeu de données d'entraînement disponible. En effet, les méthodes que nous avons décrites au chapitre 2, la génération de questions à partir de modèles et de terminologies, la génération de paraphrases par traduction pivot et l'utilisation d'embeddings contextuels pour incorporer des connaissances latentes dans le système, nous ont permis de montrer qu'il était possible d'entraîner un modèle de compréhension de la langue en l'absence de jeu de données d'entraînement. Nous avons évalué ce modèle sur un jeu de données de vie réelle et obtenu des performances suffisantes pour envisager de l'utiliser comme base pour mettre en place une première version d'un système de dialogue pour interroger le dossier patient en langue naturelle.

Nous avons également pu voir, l'importance des embeddings contextuels et des données sur lesquels ils étaient entraînés pour les tâches de compréhension de la langue. En effet, dans toutes les expériences que nous avons pu faire, les résultats obtenus avec les embeddings contextuels (ELMo, BERT) entraînés sur de larges corpus étaient meilleurs que les embeddings classiques. Sur les tâches très dépendantes du vocabulaire, comme la reconnaissance d'entités nommées, les performances étaient améliorées quand les embeddings contextuels étaient entraînés ou fine-tunés sur un corpus du domaine. En effet, les distributions des mots et leur utilisation dans les corpus de textes cliniques de DPI sont différentes du domaine général, aboutissant à des résultats inférieur lors de l'utilisation d'embeddings entraînés sur le domaine général. Il sera intéressant de tester également les récents modèles BERT entraînés sur des corpus du domaine biomédical comme BioBERT.[199] Au delà de l'apprentissage auto-supervisé des modèles tels que BERT ou ELMo, d'autres approches permettent de renforcer les représentations. Les techniques d'apprentissage semi-supervisées comme *self-training* ou *knowledge distillation*, permettent d'injecter des connaissances dans les représentations d'embeddings en utilisant de faibles quantités d'exemples annotés.[200] Transposer ces approches sur les tâches qui nous intéressent pourrait permettre d'améliorer les performances des représentations contextuelles disponibles.

De la même façon, l'intégration de connaissances expertes nous a permis d'améliorer les résultats de la reconnaissance d'entités nommées sur les médicaments. Notre modèle hybride injectant les résultats d'une pré-annotation à base de règles expertes dans un biLSTM a obtenu de meilleurs résultats que le biLSTM seul.

A travers ces deux résultats, nous pouvons voir l'importance des connaissances pré-existantes.

En effet, les textes médicaux regorgent de termes qui ne sont compréhensibles qu'à condition de disposer de la grille de lecture adéquate. Or, cette grille de lecture est une base de connaissance que le professionnel de santé se forge durant son apprentissage théorique et pratique. Cette base de connaissances implicite est complexe, et hétérogène. Reconstituer une telle base de connaissances et l'incorporer aux approches classique de machine learning est un domaine recherche à part entière mais il est envisageable dans un premier temps de s'en inspirer en créant des représentations prenant en compte diverses sources de données hétérogènes. Cette approche dite multimodale a déjà été appliquée en imagerie [201] ou dans l'apprentissage de représentations de dossiers patients [202] avec des résultats encourageants mais pourrait bénéficier aux tâches de traitementde la langue.

Sur un plan plus opérationnel, nous avons créé un nouveau corpus de données annotées avec les informations sur les médicaments. Ce corpus est le premier corpus de documents issus de DPI en français, avec ces informations. Bien que non disponible complètement librement suite aux contraintes de confidentialité, ce corpus est accessible à la communauté de recherche sous réserve d'un accord du comité scientifique et éthique de l'APHP et de règles de partage des données définies par l'APHP.

La complexité des systèmes d'information hospitaliers combinée à l'hétérogénéité des outils TAL rend souvent difficile le passage à l'échelle des projets. Nous avons voulu répondre à cette difficulté en développant l'outil PyMedExt qui simplifie l'acquisition, la transformation, l'annotation et la diffusion des textes cliniques dans le cadre de projets de TAL dans le domaine médical. Cet outil est diffusé en open-source et est conçu de manière à être modulaire et adaptable.

Enfin, tirant partie des précédentes contributions, nous avons pu mettre en pratique un projet d'annotation de données médicales à grande échelle durant la crise du coronavirus. Nous avons mis en évidence l'utilité de l'information textuelle pour répondre à des questions de santé publique dans le cadre d'une maladie émergente. Au cours de cette crise nous avons appliqué les méthodes développées durant cette thèse sur les textes cliniques des patients COVID-19 au niveau de l'ensemble de l'APHP pour extraire les informations pertinentes sur cette cohorte de patients hospitalisés. Grâce aux outils déployés, ces informations ont pu être mises à disposition, au jour le jour, des équipes de pilotage de la crise mais également des équipesde recherche. Un certain nombre de travaux utilisant ces résultats ont d'ores et déjà été publiés.[195–197,203] Extraire les données textuelles et les restructuré nous a également permis de participer, à un consortium de recherche international de recherche sur la COVID-19 [204] [1] En effet, la restructuration des informations vers des terminologies standardisées permet de s'abstraire des limites liées à la langue.

Au total, l'ensemble de nos contributions participe à rendre possible de nouvelles formes d'interactions avec le DPI. L'interrogation en langue naturelle du DPI permettra d'augmenter l'efficience et la fluidité de la recherche d'information dans le DPI tout en diminuant la charge cognitive des professionnels. Cela permettra également d'introduire des couches supplémentaires d'intelligence comme par exemple l'interprétation des résultats d'examen. Un tel système pourra, en effet, répondre à une question concernant un diagnostic en interprétant un résultat d'examen biologique pertinent (*e.g.*, répondre à une question concernant la présence d'une anémie en interprétant le taux d'hémoglobine). L'extraction d'informations dans les textes cliniques permettra de remplir automatiquement des formulaires standardisés de données structurées. Elle pourra également fournir une aide à la documentation médicale en permettant de réaliser des extractions en temps réel mettant en évidence les informations déjà saisies et pointant éventuellement vers celles qui

1. https://covidclinical.net/

manquent encore. En d'autres termes, ces avancées permettront de rendre l'utilisation des DPI plus humaine.

Annexe A

Guide d'annotation pour les données médicamenteuses dans les textes cliniques en français

A.1 Overview :

This Guideline is done to help and give standard method of annotation for medication extraction from french electronic health records. It's strongly inspired from Preliminary Annotation Guidelines of the i2b2 Medication Extraction Challenge of 2009. For each patient report provided, the goal is to extract information about all of the medications that are known to be taken by the patient or related with him. Some of the medications are provided in semi-structured (list) form, e.g., sections labeled as "medications on admission" or "medications at discharge." The final output of medications of the patient should include these ; however, the real interest is in extracting medications that are mentioned in the narratives of records.

The input of the medication annotation will be discharges summaries of Electronics Health records, pre annotated by a rule based system. The annotation will be done with "brat" [1]. Those discharges are free text.

The output created from these annotations will be a list of medications and their informations. For each listed medication, the following information needs to be annotated if missing or corrected if false in the pre annotation :

1. medication name (or drug class)
2. dosage
3. frequency
4. duration
5. mode/route of administration
6. condition
7. event
8. attributes
9. prescription patterns

1. Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou and Jun'ichi Tsujii (2012). brat : a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*.

Each entity can be annotate by attribute markers :

1. Duration and Event entity will be marked by temporal attribute(past, present, futur). If it's missing, "present" will be considered by default as temporal attribute.
2. All of the entity will be marked by certainty attribute (factual, suggested, conditional, uncertain, negated, contraindicated). If it's missing, "factual" will be considered by default as certainty attribute.
3. For all of the listed medications, the patient will be considered as experiencer by default. If it's not the case, precise it : family if it concerns patients family, "other" otherwise (notice, dosage…)

Annotations must be done even if there are spelling mistakes, unless those spelling mistakes could induce confusion.

In this document, accents and some punctuation signs have been removed from french example sentences to agree with our first study on extraction of medication informations.

A.2 Medication :

All Medications listed in discharge summary and given (present, past or future) or contraindicated to an experiencer.

A.2.1 What should be annotated ?

Drug name, generics, class of medication or substance

Medications include :

- Prescription substances :
  - Brand name medications, e.g., *doliprane*
  - Generics , e.g.,* paracetamol*
  - Ingredients, e.g., *furosemide*
  - Collective name for a group of medications, e.g., *corticoïdes* (it will be annotated as a drug class)
- Over the counter medications :
  - Brand names, e.g., *Aspirine*
  - Ingredients, e.g., *vitamine D*
  - Collective name for a group of medications, e.g, *vitamines* (it will be annotated as a drug class)
- Biological substances required or suggested by doctors
  - Ingredients in total parenteral nutrition if listed individually
  - Components of IV fluid and saline listed (including "eau minerale" and "serum physiologique")
  - Debit glucidique
- Substance therapy , e.g., *Corticothérapie* or *traitement antiretroviral* (it will be annotated as a drug class)

Medications exclude :

- Food and Water not used as treatment
- Diet
- Tobacco
- Alcohol

- Illicit drugs
- medical device , e.g., Pompe à Insuline (even if a drug is in)
- transfusion

Class include :

- "therapie"
  - oxygenotherapie
  - corticotherapie
  - antibiotherapie
  - antibiotique
- "traitement anti-" with or without "traitement" before
  - traitement antalgique
  - traitement antiretroviral
  - antihypertenseur
  - vaccination antigrippal
  - antifongique
- chemotherapy protocol
  - abvd
- others
  - vaccination contre l'hepatite b/meningocoque…
  - nutrition (if used as treatment)
    - "orale/parentérale/entérale"
      - include medication name
      - unless there is a distinction between them "nutrition orale mais pas enterale"
    - regime d'urgence
  - complement nutritionnel
- acronyme
  - np (for nutrition parenterale)
  - avk (for anti-vitamine K)
  - o2
  - vit-d

Class exclude :

"traitement" without precision

- traitement prophylactique
- traitement pour son asthme

action followed by "par"

- sedation
- (re)hydratation
- antibioprophylaxie

class included in medical device

To annotate medications, the text has to include an explicit statement indicating that the patient either took this medication, is taking the medication, is prescribed the medication, is suggested to take the medication, had side effect taking the medication or can't take it because of contraindication.

For medication suggested or uncertain, a certainty attribute"suggested"/"uncertain" must be add. For medication not taken or not given, a certainty attribute "negated" must be add (relation on a negated drug can be annotate, e.g., relation between avk and duration must be annotate for "pas d'avk pendant 2 jours"). For medication mentioned as a contraindication, a certainty

"contraindicated" must be annotate. If medications concerned other people, it must be annotated with an experiencer attribute ("family"/"other").

A.2.2 How to annotate ?

Annotate the complete noun phrase that correspond to the name of the medication, e.g., amoxicilline acide clavulanique. Annotation must be done even if there are spelling mistakes. Don't Include words such as "injectable," "creme," "nebuliseur," "solution" as part of the medication name even when they appear immediately after the medication name, e.g., selenium injectable, xylocaïne nebuliseur. Don't include numeric informations as part of the medication, e.g., renutril 500 unless it concerns a kind of the substance, e.g., iodure 131

Pronouns that refer to a drug, mustn't be included, but its attributes are related with the element referred.

Examples :

- *amlor : 10 mg le matin*
  - drug : *amlor*
- *ialuset plus creme*
  - drug : *ialuset plus*
  - (creme not included in drug name)
- *sevrage de l oxygenotherapie en fevrier 2013*
  - class : *oxygenotherapie*
- *grand-mere maternelle : diabete de type 2 a 61 ans, sans surpoids, traitee par insuline*
  - drug : *insuline*
    - experiencer : family

Each co reference of a medication (class or drug name) or its generics, including with spelling istakes, in the same sentence, must be annotated.

- *doliprane 1 dose poids*4/ jour si douleurs (paracetamol 1 boite)*
  - drug : *doliprane*
  - drug : *paracetamol*
- *(matin : 9-12 ui novorapid(1), 20 ui levemir(1), dans les 4 zones. (gouter : 5-7 ui novorapid(2). (soir : 6-8 ui novorapid(3), 15 ui levemir(2), dans les 4 zones.*
  - drug : *novorapid(1)*
  - drug : *levemir(1)*
  - drug : *novorapid(2)*
  - drug : *novorapid(3)*
  - drug : *levemir(2)*

If a drug is written as medication and a class in the same sentence, annotate both (as a class and as a drug). Medication association with class in the same sentence results of one annotation "drug" by medication :

- *relais par avk au cours de l'hospitalisation (coumadine)*
  - drug : *coumadine*
    - is_part_of : *avk*
  - class : *avk*
- *un traitement antiretroviral a ete debute (truvada, reyataz et norvir avec une charge virale…)*
  - class : *traitement antiretroviral*

    - drug : *truvada*
    - drug : *reyataz*
    - drug : *norvir*

Medication enumeration sharing a word must be annotated together :

- *vitamine C , D , A , and E*
    - drug : *vitamine C , D , A , and E*

But :

- *une dose de vitamine C et vitamine D*
    - drug : *vitamine C*
    - drug : *vitamine D*

Annotate drugs name even if their attributes are negated

- *pas de necessite de doublement des doses d hydrocortisone*
    - drug : *hydrocortisone*
    - don't annotate "doublement des doses"

A.3 Dosage :

The amount of a single medication used in each administration, e.g., *un comprimé, une dose, 30 mg.*

A.3.1 What should be annotated ?

The numeric and/or the textual information that mark the amount and the unit of administration of a medication used in a single administration. Annotate relation with the drug concerned by a link from dosage to drug name

Includes (not exhaustive) :

- 1 cp
- un comprimé
- 0.4 mg
- 0.5 m.g.
- 100 mg/kg
- une dose kilo
- 100 mg x 2 comprimés
- 500 mg : 2 gelules
- 4000 ui / 0,4
- 1 sachet
- 2 cuillère-mesure
- deux bouffées
- 3 bolus
- double dose
- [renutril] 500

Exclude :

if dose is negated and drug is given, e.g., don't annotate dose ,

- don't annotate "doublement des doses" for "pas de necessite de doublement des doses d hydrocortisone"

Cumulative dosages (because too much variability in the meaning) :

- 3 boites

A.3.2 How to annotate ?

Annotate all mentioned dosages of all medications present in the discharge summary and their relation with it even if it is part of the medication name.

- *speciafoldine 5mg 10 jours par mois*
  - Dosage : *5mg*
- *oracilline 500mui x 2 par jour.*
  - Dosage : *500mui*
- *depakine 500 x 3 par jour*
  - Dosage : *500*
- *vitabact 0,05 % : x4/jour dans chaque oeil pendant 10 jours*
  - Dosage : *0,05 %*

Annotate all the partial dosage as "Dosage"

- *hydrea 500mg un jour sur 2, 1000mg un jour sur 2.*
  - Dosage : *500mg*
  - Dosage : *1000mg*
- *hydrocortisone : 7,5 mg le matin, 5 mg le soir (12,5 mg/m²/jour)*
  - Dosage : *7,5 mg*
  - Dosage : *5 mg*
  - Dosage : *12,5mg/m²*

Annotate different ways of referring to the same dosage in separate entries :

- *sandostatine : 100µg/8h en sc soit 50µg/kg/j*
  - Dosage : *100µg*
  - Dosage : *50µg/kg*

Annotate immediately adjacent part of a dosage in seperate entry :

- *seretide 50 deux bouffeesx2/j*
  - Dosage : *50 deux bouffees*
- *singulair 1 sachet de 4mg/jour,*
  - Dosage : *1 sachet de 4mg*

Annotate a range of dosage as one entry. In this example, there is multiple dosage for the same drug but in different sentence :

- *matin : 5 a 8 ui novorapid. midi : 5 a 8 ui novorapid. gouter : 3 a 4 1/2 ui novorapid. soir : 3 a 6 ui novorapid.*
  - Dosage : *5 a 8 ui*
  - Dosage : *5 a 8 ui*
  - Dosage : *3 a 4 1/2 ui*
  - Dosage : *3 a 6 ui*

Annotate only one pattern (ordonnance prescription) for all drug when dosage concerned both :

- *doliprane et ibuprofene, 1 comprime toutes les 6 heures chacun*
  - Dosage : *1 comprime*
  - Ordonnance_prescription_start : *doliprane et ibuprofene, 1 comprime*

A.4 Frequency

Terms, phrases, or abbreviations that describe how often each dose of the medication should be taken.

A.4.1 What should be annotated ?

Any expression that indicates the frequency of administering a single dose of a medication should be annotated.

Includes :

Frequency :
- par jour
- toutes les 8 heures
- toutes les semaine
- /jour
- /j
- /24 heures
- par mois
- tous les soirs
- x 3 par jour
- jour (if preceded by dosage)
- 1 - 1 - 1
- 850 - 1000 - 1000
- 19/03, 25/03 et 01/04/2016
- a J3, J5 et J7

Temporal phrases which specify when a medication should be taken (These tend to be prepositional phrases. Preposition should be included in the extracted information) :
- quotidiennement
- quotidien
- mensuel
- après/avant manger
- a 4 heures
- avant chaque repas

A.4.2 How to annotate ?

Apply the same basic principles that you use for tagging dose. Annotate each frequency even if repeated in the same sentence
- doliprane 1 dose poids*4/ jour si douleurs
  - Frequency : *4/ jour
- sectral : 48 mg matin et soir
  - Frequency : matin et soir

Annotate immediately adjacent part of a frequency as one entry :
- singulair a chaque bolus : 15g a 7h et 16h30
- Frequency : a 7h et 16h30

If frequency is segmented and concerns same entity, annotate the most informative part :
- speciafoldine : 1 comprime par jour, 10 jours par mois.
  - Frequency : 10 jours par mois

A.5 Duration

A elapsed time expression that indicate for how long the medication is to be administered. Such expressions are often noun phrases, prepositional phrases, or clauses.

A.5.1 What should be annotated ?

Expressions that describe the total time period for which the medication should be taken at a given dose. In case of medications that are stopped, the duration indicates for how long the medication has been stopped.

Includes :

Time expressions :
- [pendant] 10 jours
- [pour] un mois
- [durant] 2 semaines
- tant que nécessaire
- [sur] 3h

Excludes :

Time expressions that indicate when each dose should be taken. Include these under frequency.
- a prendre pendant une activité physique
  - "pendant une activite physique" is not a duration

Time expression of starting or stopping a medication :
- dans 10 jours
- depuis 10 jours

Cumulative dosages (because too much variability in the meaning) :
- 3 boites

A.5.2 How to annotate ?

Follow the same basic principles as for annotating frequency. Don't include complete prepositional. Duration must be annotated with temporal attribute. if missing, "present" will be considered by default as temporal attribute.
- vitabact 0,05 % : x4/jour dans chaque oeil pendant 10 jours
  - Duration : 10 jours
    - temporality : present (by default)
- ciflox 500 mg par 24 heures pour une duree totale de 3 semaines
  - Duration : 3 semaines
    - temporality : present (by default)
- a donc beneficie de sa 2ieme perfusion de remicade (200mg) sur 3h
- Duration : 3h
  - temporality : present (by default)
- a ete traite pendant 2 ans par remicade
  - Duration : 2 ans
    - temporality : past

A.6 Mode of administration

Describes the method for administering the medication.

A.6.1 What should be annotated ?

Text that expresses mode/route of administration, even when it is expressed as part of the medication name or the dosage.

Includes :

per os
- intraveineux
- topique
- sublingual
- cutanee
- sous cutanée
- intramusculaire
- perfusion
- creme
- solution buvable
- ophtalmique
- Abbreviations of the above

A.6.2 How to annotate ?

Follow the same basic principles as for annotating duration. If route apply to multiple medication, add a relation for each. Multiple route can be related to one drug name
- ventoline spray : 2 bouffees x4 par jour pendant 4 jours au babyhaler.
- hyperhydratation par voie intraveineuse
  - route : intraveineuse
- necessitant un traitement par kayexalate et aerosol de ventoline
  - route : aerosol

Changes in mode of administration of a drug should be included as separate entries.
- traitement pendant 5 jours par clamoxyl iv puis relais per os
  - route : iv
  - route : per os

Different ways of referring to the same mode of administration should be included in separate entries.
- nebulisation de ventoline toutes les 6 heures puis relais par chambre d inhalation (baby-haler) le 06/02/2012
  - route : nebulisation
  - route : chambre d inhalation
  - route : baby-haler

Cases where one mode applies to multiple medications need to be handled properly.
- relais par targocid puis orbenine iv jusqu au 24/03/2013
  - route : iv

A.7 Condition

Expressions that indicate condition for which the medication is to be given. Such expressions are often conditional proposal and start with a conditional expression such as "si," "en cas de," "en fonction de"…

A.7.1 What should be annotated ?

Condition for which the medication is to be given.

Includes :
- en cas de fievre
- si besoin
- si veut
- en fonction des ASAT

A.7.2 How to annotate ?

Always annotate the most informative base adjective phrase or the longest base noun phrase as the condition for the medication. Longest base noun phrase has the form (det* adj* N+ adj*). Longest adjective phrase often occurs as (adj+). Do not include complex phrases, do not include coordinated phrases. Instead, extract from these phrases the base phrase, even when this means you will end up with multiple conditions. A condition can be related to drug name or an event.

A certainty attribute"conditional" must not be add on the entity concerned by the condition.
- codenfan une dose/poids si besoin maximum 3x par jour
  - drug : codenfan
  - condition : besoin
    - is_condition : codenfan

If there are different conditions mentioned for the same medication then include one entry per condition. Add relation with entity for each. In cases where multiple medications are given with the same condition, list the condition with all of the medications and add relation for each.
- il a ete explique aux parents d utiliser l oxygene en cas d inconfort, de paleur ou de gene respiratoire et non en fonction d un chiffre de saturation
  - drug : oxygene
  - condition : inconfort
  - condition : paleur
  - condition : gene respiratoire
  - condition : chiffre de saturation
    - certainty : negated

If a condition is composed of multiple sub-conditions (separated by "et"), annotate them together with one entry.
- melatonine 2mg : 1 gelule au coucher si agitation et probleme d endormissement
  - condition : si agitation et probleme d endormissement

Different ways of referring to the same condition for medication should be treated as separate conditions. Add relation "is_equiv" between them and to the related entity from the closest one.
- en cas d'anemie aregenerative (hemolyse non mecanique) augmenter les corticoides
  - condition : anemie aregenerative
  - condition : hemolyse non mecanique

- increase : augmenter
  - Arg : corticoides

A.8 8.Events

Information on whether the medication is started, stopped, continued, increase or decrease at a defined time. This information is usually expressed in the main verb of the sentence or by a date. Annotate the event indicated by the most precise date, or, if not possible, by the main word related to the medication.

A.8.1 What should be annotated ?

A date of starting, stopping, continuing, increasing or decreasing a medication. If missing, annotate the main word (verb, noun…) highlighting the event such as “mise en route,” “début,” “poursuite,” “relais”…

A.8.2 Possible relations from an event

“Arg” of a medication : links the event to the related medication affected by the event. To an entity. a medication could have multiple events related.

A.8.3 How to annotate ?

Choose from possible values : start, stop, continue, start-stop, increase or decrease.

- Start : main date or word of the medication beginning
- Stop : main date or word of the medication stop
- Start-Stop : unique intake of a medication
- Increase : increase of dosage of a medication already taken
- Decrease : decrease of dosage of a medication already taken
- Continue : main date or word of the medication pursuit or with no obvious change of dosage

By default, “present” is the temporal attribute of Events. It can be “past,” “present” or “future.” It must be defined according to if the event is before, during or after current hospitalization.

If there are two Events on the same expression (even if both of them are the same, for example 2 start events), you should annotate the expression twice with an event “Start”

A.8.4 Example :

- Pas de modification de la corticothérapie
  - continue : Pas de modification
    - Arg : corticothérapie
- meningocoque a + c : 11/07.
  - start-stop : 11/07
    - Arg : meningocoque a + c
    - Temporality : past
- antibiotherapie debutee lors de la chirurgie, a arrete a j5
  - start : debuter

        - Arg : antibiotherapie
    - stop : a j5
        - Arg : antibiotherapie
- arret du nubain le 14/12/2010
    - stop : 14/12/2010
        - Arg : nubain
        - Temporality : past
- augmentation des doses de morphine
    - increase : augmentation
        - No date, so main word is "augmentation"
        - Arg : morphine
        - Temporality : present (by default)
- poursuite de l'hydrea
    - continue : poursuite
        - Arg : hydrea
- janvier 2006 : nouveau syndrome thoracique aigu, mise sous hydrea.
    - start : janvier 2006
        - Arg : hydrea
        - Temporality : past
- compte-rendu d hospitalisation de jour du 27/12/2012 pour sa 16ieme perfusion de remicade
    - start-stop : 27/12/2012
        - Arg : remicade
        - Temporality : present

Switching one medication for another includes two events on the same expression. One medication is stopped and another one is started.

- relais du traitement avk pour un traitement par heparine en sous cutanee dans la phase aigue
    - stop : relais
        - Arg : avk
        - Temporality : present
    - start : relais
        - Arg : heparine
        - Temporality : present

If there are two events for an entity, include two separate entries. Make each entry as specific and complete as possible. If there are multiple medications for one event, include separate entries for each. Add a relation for each.

- la pancytopenie s est compliquee apres la chimiotherapie d un sepsis a escherichia coli resistant a la tazocilline (tazocilline* depuis le 6 septembre 2010) traite par fortum a partir du 15 septembre 2010
    - start : 6 septembre 2010
        - Arg : tazocilline
        - Temporalité : present
    - stop : 15 septembre 2010
        - Arg : tazocilline
        - Temporalité : present
    - start : 15 septembre 2010
        - Arg : fortum

    - Temporalité : present
- traitement par endoxan avant de debuter un traitement par mabthera fludarabine endoxan etant donne la lymphocytose majeure et la presence d anemie hemolytique
    - start : debuter
        - Arg : mabthera
    - start : debuter
        - Arg : fludarabine
    - start : debuter
        - Arg : endoxan
- debut du traitement par ambisome le 29 mars 2014 a 3 mg/kg jusqu au 2 avril puis 5 mg/kg jusqu au 7 avril, puis 7,5 mg/kg jusqu au 30 avril
    - start : 29 mars 2014
        - Arg : ambisome
    - increase : 2 avril
        - Arg : ambisome
    - increase : 7 avril
        - Arg : ambisome
    - stop : 30 avril
        - Arg : ambisome

A.9 Attributes

Information that indicates when events are to take place, and whether they are factual, suggested, conditional or uncertain

A.9.1 Temporal Attributes

Information about whether the medication was administered in the past, is being administered currently, or will be administered in the future, to the extent that this information is expressed in the tense of the verbs and auxiliary verbs used to express events. One temporal attributes for each event.

How to mark ?

Events and duration can be marked.

Choose from possible values for each event : past, present, future. The default temporal attributes is “present.”

- Past : The event occurred before current hospitalization.
- Present : The event occurred during current hospitalization.
- Future : The event occurr after current hospitalization.

See Duration and Events for examples

A.9.2 Certainty attributes

Information on whether the event occurs. Certainty can be expressed by uncertainty words, e.g., “suggested,” or via modals, e.g., “should” indicates suggestion.

How to mark ?

Choose from possible values : conditional, suggestion, factual, uncertain or negated. The default Certainty attributes is "factual"

- Conditional : The entity occurs only under certain conditions as mentioned in the text.
- Suggestion : The entity is/was suggested and it is not dependent on a condition.
- Factual : The entity is not marked as conditional or suggestion, it is factual. This is the default value for certainty.
- Uncertain : The entity does/did not occur for sure.
- Negated : The entity does/did not occur, e.g., a medication is not given
- Contraindicated : the medication is mentioned as a contraindication

See previous chapter for examples

A.10 Prescription Pattern

All Medications and their attribute listed in discharge summary and given (present, past or future) or contraindicated to an experiencer.

A.10.1 What should be annotated ?

There is two kind of pattern : medication prescription (medication blob) and ordonnance prescription (ordonnance blob). A drug (and its equivalent) and all of its attributes must be annoted such as medication prescription. A sentence with multiple drugs and their attribute must be annoted such as ordonnance prescription. if an attributes is related to multiple drugs, it must not be in a medication prescription but only in ordonnance prescription.

A.10.2 How to annotate ?

From the first word (medication, attributes or event) until the last. Event type must be applied on the Prescription patterns related. if a drug have no attributes, do not do a prescription pattern.

Examples :

- doliprane 1 dose poids*4/ jour si douleurs (paracetamol 1 boite)
  - drug : doliprane
  - drug : paracetamol
  - dose : 1 dose poids
  - frequency : 4/jour
  - condition : douleurs
  - medication_blob : doliprane 1 dose poids*4/ jour si douleurs (paracetamol

if an attribute or event is related to multiple drugs, include it only in the "ordonnance_blob" pattern which will include the drugs. Here "toujours" is related to the ordonnance.

- je ne modifie pas son traitement, soit toujours lasilix 20 mg/j, atacand 8 mg, ezetrol , calciparat 1 g, allopurinol 300 mg et crestor 5.
  - event : continue
    - Arg : toujours
  - drug : lasilix
  - dose : 20 mg
  - frequency : /j

- medication_blob : lasilix 20 mg/j
- drug : atacand
- dose : 8 mg
- medication_blob : latacand 8 mg
- drug : ezetrol
- drug : calciparat
- dose : 1 g
- medication_blob : calciparat 1 g
- drug : allopurinol
- dose : 300 mg
- medication_blob : allopurinol 300 mg
- drug : crestor
- dose : 5
- medication_blob : crestor 5
- ordonnance_blob : toujours lasilix 20 mg/j, atacand 8 mg, ezetrol , calciparat 1 g, allopurinol 300 mg et crestor 5

Informations Offset

The informations' offsets will be generated automatically by "brat." It starts by 0 for the first character of the discharge summary. Each character, even space, counts as one character. Return to the line count as 2 characters.

Each entry consist of one annotation (entity or relation) and will be printed on its own line with its offset. All annotations follow the same basic structure : each annotation is given an ID that appears first on the line, separated from the rest of the annotation by a single TAB character. The rest of the structure varies by annotation type.

For entity, the line follows those example : T# "type of entity" start stop "Entity"

T1 drugs 1760 1770 solumedrol

T2 dose 1771 1777 180 mg

Entities annotated can be fragmented in multiple part (in this document, it is represented as [...] but not in the offset). The line follows the format : T# "type of entity" start1 stop1 ;start2 stop2 "Entity"

For relation between 2 arguments , the line follows the format ; R# "type of relation" Arg1:fromID Arg2:toID

R1 is_dosage Arg1:T2 Arg2:T1

Each event have at list 2 lines : one for the highlighting word (like an entity) and one for each "Arg" relation : E# "Event type":"Event ID" Arg:toID

T100 start 3915 3927 introduction E1 start:T100 Arg:T99

Attribute follows the format : A# "Attribute class" "ID of Entity marked" "Attribute name"

A1 certainty T109 uncertain

Annexe B

Table des questions “vie réelle”

TABLE B.1 : Questions 'vie réelle'

question	entity_type	result_type	result_time	time_constraint	interpretation
la clairance de la créatinine est-elle normale ?	result_uniq	value	last	none	normality
la croissance de l'enfant est-elle normale ?	diagnostic	value	all	none	normality
y a t-il une insuffisance rénale ?	diagnostic	presence	last	none	NA
Y t-il une hémolyse ?	diagnostic	presence	last	none	NA
existe t-il un syndrome inflammatoire ?	diagnostic	presence	last	none	NA
y a t-il une infection urinaire ?	diagnostic	presence	last	none	NA
à quand remonte la dernière échographie rénale ?	result_uniq	date	last	none	value
quelle est la date et le résultat de la dernière biopsie rénale ?	result_uniq	date,result	last	none	value
Comment le taux de neutrophiles a-t-il évolué entre telle et telle date ?	result_uniq	evolution	all	range	value
Comment le taux de neutrophiles a-t-il évolué depuis l'introduction du traitement XXX ?	result_uniq	evolution	all	range	value
Quelle est la date de sortie d'aplasie post allo greffe de cellules souche hématopoïétiques	diagnostic	date	last	none	value
évolution de la CRP/procalcitonine depuis 7 jours	result_uniqs,result_uniqs	evolution	all	range	value
identification d'un germe sur 1 mois ?	diagnostic	value	all	range	presence
identification d'une examen direct positif sur 7 jours ?	diagnostic	value	all	range	presence
liste des prélèvements envoyés en microbiologie pour analyse	lab_order	presence	all	range	NA
Suivi des leucocytes/PNN/lymphocytes/monocytes/eosinophiles sur 1 semaine	result_uniqs,result_uniqs,result_uniqs,result_uniqs,result_uniqs	evolution	all	range	value
Suivi de l'hémoglobine sur 3 jours	result_uniq	evolution	all	range	value
Suivi PCR EBV sur 1 an	result_uniq	evolution	all	range	value
Suivi PCR CMV sur 2 ans	result_uniq	evolution	all	range	value
Suivi clonalité (T/B) sur 3 mois	result_uniq	evolution	all	range	value
Suivi BCR-ABL sur 15 jours	result_uniq	evolution	all	range	value
Suivi des lymphocytes et % de la population tumorale sur 6 mois	result_uniqs,result_uniqs	evolution	all	range	value
Dernière hémoculture positive ?	result_uniq	date	last	none	positivity
Tableau des résultats positifs bactériologiques	diagnostic	value	all	none	positivity
Tableau des résultats de cytométrie de flux fait au CEDI	result_uniq	value	all	none	value
Tableau des résultats de génétique	result_uniq	value	all	none	value
quel est le résultat du bilan hormonal lors d'une hypoglycémie ?	result_group	value	last	date	value
Quel sont les ammoniémies depuis ce matin ?	result_uniq	value	all	range	value
Quelles sont les résultats du cycle glycémie lactate depuis hier ?	result_group	value	all	range	value
Quelles sont les résultats de génétique de ce patient ?	result_uniq	value	all	none	value

TABLE B.1 : Questions 'vie réelle' *(continued)*

question	entity_type	result_type	result_time	time_constraint	interpretation
Quels sont les résultats de son test au glucagon ?	result_uniq	value	last	none	value
Quels sont ses CPK ? Sont ils en train de diminuer significativement ?	result_uniq	value,evolution	last,all	none	value,low
L'hémoculture sur son KTC est elle positive ?	result_uniq	value	last	none	positivity
Comment était sa dernière chromatographie des acides aminés plasmatiques ?	result_group	value	last	none	value
Quel est son trou anionique ?	diagnostic	value	last	none	value
acuité visuelle	result_uniq	value	last	none	value
pression intraoculaire	result_uniq	value	last	none	value
épaisseur maculaire	result_uniq	value	last	none	value
quelle était la dernière valeur de l'hémoglobine	result_uniq	value	last	none	value
quel était le dernier ECBU	result_uniq	value	last	none	value
quel est le dernier compte rendu de scanner cérébral	result_uniq	value	last	none	value
quel est le dernier compte rendu d'hospitalisation	result_uniq	value	last	none	value
Quel est le résultat de la dernière ponction lombaire ?	result_group	value	last	none	value
A t il déjà eu une ponction lombaire ?	result_group	presence	all	none	NA
Quel est le résultat des PCR virales dans le LCR ?	result_group	value	last	none	value
A t il un syndrome inflammatoire ?	diagnostic	presence	last	none	NA
A t il eu une recherche de virus dans le nez ?	result_uniq	presence	last	none	NA
A t il eu une recherche de virus dans les selles	result_uniq	presence	last	none	NA
A t il eu une recherche de mycoplasme	result_uniq	presence	last	none	NA
A t il eu un bilan infectieux de maladie neuroinflammatoire complet ? Tous les resultats sont ils disponibles ?	result_group	presence	last	none	completeness
A t il déjà eu des chromatographies du sang ? Des urines ? Du lcr ? Avec quels résultats ?	result_group	presence,value	last	none	value
Quel est le résultat des neurotransmetteurs dans la ponction lombaire	result_uniq	value	last	none	value
quel est le résultat de son dernier INR ?	result_uniq	value	last	none	value
comment l'INR a évolué depuis un mois ?	result_uniq	evolution	all	range	value
comment l'INR a évolué depuis 6 mois ?	result_uniq	evolution	all	range	value
quel est le résultat de son dernier NTproBNP ?	result_uniq	value	last	none	value
comment le NTproBNP a évolué depuis 6 mois ?	result_uniq	evolution	all	range	value
quel est le résultat du dernier dosage de plaquettes ?	result_uniq	value	last	none	value
quel est le résultat du dernier antiXa ?	result_uniq	value	last	none	value
comment l'antiXa a évolué depuis 5 jours ?	result_uniq	evolution	all	range	value
quel est le résultat du dernier dosage de troponine ?	result_uniq	value	last	none	value
comment la troponine a évolué depuis 7 jours ?	result_uniq	evolution	all	range	value
quel est le résultat du dernier dosage de transaminases ?	result_uniq	value	last	none	value
comment les transaminases a évolué depuis 7 jours ?	result_uniq	evolution	all	range	value
quel est le résultat de sa dernière TSH ?	result_uniq	value	last	none	value

TABLE B.1 : Questions 'vie réelle' *(continued)*

question	entity_type	result_type	result_time	time_constraint	interpretation
quel est le résultat de sa dernière hémoglobine ?	result_uniq	value	last	none	value
comment la CRP a évolué depuis 7 jours ?	result_uniq	evolution	all	range	value
de quand date la dernière hémoculture positive ?	result_uniq	date	last	none	positivity
quel est le résultat du dernier dosage de tacrolimus ?	result_uniq	value	last	none	value
quel est le résultat du dernier dosage de ciclosporine ?	result_uniq	value	last	none	value
quel est le résultat du dernier dosage de cellcept ?	result_uniq	value	last	none	value
A t'il eu une recherche de BMR ou BSLE ?	result_group	presence	last	none	NA
Quelle est la date de recherche de BMR ou BSLE ?	result_group	date	last	none	NA
Quelle est l'évolution de la bilirubine conjuguée ?	result_uniq	evolution	all	range	value
Quelle est l'évolution des ASAT ?	result_uniq	evolution	all	range	value
Quelle est l'évolution des ALAT ?	result_uniq	evolution	all	range	value
Quelle est l'évolution des GGT ?	result_uniq	evolution	all	range	value
Quelle est la date du dernier bilan biologique ?	result_group	date	last	none	value
Quel est le résultat de l'électrophorèse des protéines ?	result_group	value	last	none	value
A t'il plus de 50000 plaquettes au dernier bilan ?	result_uniq	value	last	none	comparison_to_value
Quel est le nombre d'éosinophiles ?	result_uniq	value	last	none	value
Y'a t'il une hypoalbuminémie ?	diagnostic	presence	last	none	NA
Quelle est l'évolution de l' antithrombine 3 ?	result_uniq	evolution	all	none	value
Quelle est l'évolution du TP ?	result_uniq	evolution	all	none	value
quel est le résultat bactériologique de la biopsie du 4/02/17 ?	result_uniq	value	last	date	value
quelle est le résultat de la dernière NFS ?	result_group	value	last	none	value
Quel est le dernier bilan bio réalisé ?	result_group	presence	last	none	NA
Le patient a t il une carte de groupe sanguin ?	NA	NA	NA	NA	NA
Les RAI ont elles été faites et sont elles à jour ?	result_uniq	presence, date	last	none	, comparison_to_value
Y a t il un bilan d'hémostase connu pour ce patient ?	result_group	presence	last	none	NA
Quels sont les résultats du bilan d'hémostase ?	result_group	value	last	none	value
Quelle est la cinétique de la numération plaquettaire sur les 6 derniers mois ?	result_uniq	evolution	all	range	value
quels sont les résultats du dernier bilan	result_group	value	last	none	value
quel est le chiffre maximal de CRP/PCT	result_group	value	last	none	maximum
quel est le résultat de tous les items de bio disponibles dans Stare	result_group	value	all	none	value
le patient est il porteur de BMR ?	result_group	value	last	none	positivity
comment a évolué l'antigénemie aspergillaire	result_uniq	evolution	all	range	value
comment a évolué le taux de beta D glucane	result_uniq	evolution	all	range	value
de quand date le dernier ECBU	result_uniq	date	last	none	value
quel est le résutlat du dernier prélèvemetn vaginal	result_uniq	value	last	none	value
comment est la clearance de la creatinine	result_uniq	value	last	none	value

TABLE B.1 : Questions 'vie réelle' *(continued)*

question	entity_type	result_type	result_time	time_constraint	interpretation
Quel est le dernier résultat des transaminases ? des GammaGT ? de la bilirubine totale et conjuguée ? de l'albuminémie ? des gammaglobulines ? du TP ? des cofacteurs de la coagulation	result_group,result_group,result_group,result_group,result_group,result_group,result_group	value,value,value,value,value,value,value	last	none	value
Quelle est l'évolution des transaminases ?cdes GammaGT, de la bilirubine totale et conjuguée ?	result_group,result_group,result_group	evolution,evolution,evolution	all	none	value
Est-il immunisé vis à vis de l'hépatite A ?	result_uniq	value	last	none	positivity
Est-il immunisée vis à vis de l'hépatite B ?	result_uniq	value	last	none	positivity
Quelle est le résultat du dernier dosage de l'alphafoetoprotéine sérique ?	result_uniq	value	last	none	value
De quand date le dernier dosage de l'alaphafoetoprotéine ?	result_uniq	date	last	none	value
De quand date le dernier contrôle des IgG totaux anti HV et des ac anti HBs ?	result_uniq,result_uniq	date	last	none	value
Y a t'il un résultat de PCR EBV ? CMV ? HHV6 ?	result_uniq,result_uniq,result_uniq	presence,presence,presence	last	none	NA
Y a t'il une hypergammaglobulinémie ?	diagnostic	presence	last	none	NA
Quel est le résultat du dernier dosage de 25OH-Vit D ?	result_uniq	value	last	none	value
De quand date le dernier dosage de 25OH-Vit D ?	reuslt_uniq	date	last	none	value
Quelle est la clearance de la créatinine ?	diagnostic	value	last	none	value
Y a t'il des signes d'insuffisance rénale ?	diagnostic	presence	last	none	NA
Y a t'il des signes d'infection ?	diagnostic	presence	last	none	NA
Quel est le dernier taux d'albumine ?	result_uniq	value	last	none	value
Existe-t-il une protéinurie significative ?	diagnostic	presence	last	none	NA
Depuis quand une cytolyse est-elle présente ?	diagnostic	date	all	range	first
Depuis quand une cholestase est-elle présente ?	diagnostic	date	all	range	first
Existe-t-il une carence martiale ?	diagnostic	presence	last	none	NA
Quel est le dernier taux d'immunoglobulines ?	result_uniq	value	last	none	value
evolution de l 'hemoglobine , du taux de CD4 , du taux des GB, des lymphocytes sur les derniers mois ?	result_uniq, result_uniq, result_uniq, result_uniq	evolution	all	range	value
Y a t il une documentation microbiologique ?	result_group	value	last	none	positivity
est ce que les examens retrouvent un germe ?	result_group	value	last	none	positivity
le dernier bilan MST ?	result_group	value	last	none	value
la dernière charge virale	result_uniq	value	last	none	value
le dernier bilan mycologique	result_group	value	last	none	value
le dernier dosage en IgG	result_uniq	value	last	none	value

TABLE B.1 : Questions 'vie réelle' *(continued)*

question	entity_type	result_type	result_time	time_constraint	interpretation
Quel est le dernier résultat des anticorps anti transglutaminases ?	result_uniq	value	last	none	value
Quel est le dernier résultat de la calprotectine fécale ?	result_uniq	value	last	none	value
Quel est le résultat de la recherche d'helicobacter pylori ?	result_uniq	value	last	none	positivity
Quel est le résultat d'anapath de la dernière endoscopie ?	result_uniq	value	last	none	value
Quel est le résultat d'anapath de la dernière FOGD ?	result_uniq	value	last	none	value
Quel est le résultat d'anapath de la dernière fibro ?	result_uniq	value	last	none	value
Quel est le résultat de bacterio de la dernière endoscopie ?	result_uniq	value	last	none	value
Quel est le résultat de bacterio de la dernière FOGD ?	result_uniq	value	last	none	value
Quel est le résultat de bacterio de la dernière fibro ?	result_uniq	value	last	none	value
Quels sont les résultats des marqueurs fongiques ?	result_group	value	last	none	value
Quelle est l'évolution du syndrome inflammatoire ?	diagnostic	evolution	all	none	value
Quel est le résultat de la ponction lombaire ?	result_group	value	last	none	value
Quels sont les résultats du lavage broncho alvéolaire ?	result_group	value	last	none	value
Quelle est lévolution de la fonction rénale	result_group	evolution	all	none	value
Quelle est l'évolution du bilan hépatique ?	result_group	evolution	all	none	value
Quel est le dernier bilan immuno virologique du VIH ?	result_group	value	last	none	value
Le résultat du dernier génotypage de résistance ?	result_uniq	value	last	none	value
quelle est la ferritinémie ?	result_uniq	value	last	none	value
Quelle est la calcémie ?	result_uniq	value	last	none	value
A combien est la CRP ?	result_uniq	value	last	none	value
quelle est son dernier traitement ?	result_group	value	last	none	value
le patient reçoit-il du cellcept ?	result_uniq	presence	last	none	value
quelle est la cinétique des BHCG ?	result_uniq	evolution	all	none	value
quel est le résultat de l'amniocentèse ?	result_uniq	value	last	none	value
Quel est le résultat de la CGH / caryotype ?	result_uniq	value	last	none	value
quel est le groupe sanguin ?	result_uniq	value	last	none	value
la patiente est elle immunisée pour la toxoplasmose ?	result_uniq	value	last	none	positivity
la patiente est elle immunisée pour le CMV ?	result_uniq	value	last	none	positivity
Mr X avait-il déjà une thrombopénie lors de sa dernière hospitalisation ?	diagnostic	presence	last	range	NA
Peux-tu me donner les résultats des 4 dernières créat' de Mr X ?	result_uniq	value	last	X	value
Quelle est la vitesse d'augmentation des PSA de Mr X au cours des 4 dernières hospitalisations ?	result_uniq	evolution	all	range	value
La CRP de Mme Y à t'elle diminuée ?	result_uniq	evolution	all	none	low
Le dernier bilan de Mme Y est-il compatible avec une anémie hémolitique auto-immune ?	diagnostic	presence	last	none	NA
Des carences ont-elles été mises en évidence dans les derniers bilan de Mme Y ?	result_group	value	last	none	low

TABLE B.1 : Questions 'vie réelle' *(continued)*

question	entity_type	result_type	result_time	time_constraint	interpretation
Peux-tu me donner la date et la valeur du dernier taux de plaquette normal ?	diagnostic	date,value	last	none	value
Peux-tu me donner la date et la valeur de la première CRP de Mme Y ?	result_uniq	date,value	first	none	value
A-t-on doser des marqueurs tumoraux pour Mme Y ?	result_group	presence	last	none	NA
Le dernier bilan de Mr X est-il compatible avec une déshydratation extracellulaire ?	diagnostic	presence	last	none	NA
Quelle est la valeur de la natrémie corrigée de Mr X ?	diagnostic	value	last	none	value
Quelle est la valeur du pic de glycémie de Mme Y au cours de son hospitalisation ?	result_uniq	value	all	range	maximum
Quelle est la moyenne des hémoglobines glyquées de Mr X pour cette année ?	result_uniq	value	all	range	mean
Quelle est la valeur moyenne de la glycémie pour chacune des 4 dernières hospitalisations de Mr X ?	result_uniq	value	all	range,range,range,range	mean
Les marqueurs tumoraux de Mr X sont-ils stables au cours des 6 derniers mois ?	result_group	evolution	all	range	value
L'insuffisance rénale de Mr X est-elle stable cette année ?	diagnostic	evolution	all	range	value
Peux-tu m'afficher le graph' de l'urée et de la creat' au cours de cette année pour Mme Y ?	result_uniq,result_uniq	evolution	all	range	value
Peux tu me donner la valeur moyenne du VGM au cours de la pénultième hospitalisation pour Mme Y ?	result_uniq	value	all	range	mean
Peux tu me donner le taux d'IgE au cours de l'avant avant dernière hospitalisation pour Mme Y ?	result_uniq	value	last	range	value
quelle est la date de la dernière intervention chirurgicale ?	result_uniq	date	last	none	value
de quand date la dernière échographie abdominale ?	result_uniq	date	last	none	value

Annexe C

Questions “vie réelles” en représentation hierarchique

TABLE C.1 : Questions 'vie réelle'

question
[IN:INTERPRETATION_NORMAL la [SL:ENTITY clairance de la créatinine] est-elle normale ?]
[IN:INTERPRETATION_NORMAL la [SL:ENTITY croissance de l' enfant] est-elle normale ?]
[IN:INTERPRETATION_PRESENCE y a t-il une [SL:ENTITY insuffisance rénale] ?]
[IN:INTERPRETATION_PRESENCE y t-il une [SL:ENTITY hémolyse] ?]
[IN:INTERPRETATION_PRESENCE existe t-il un [SL:ENTITY syndrome inflammatoire] ?]
[IN:INTERPRETATION_PRESENCE y a t-il une [SL:ENTITY infection urinaire] ?]
[IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TYPE_DATE à quand remonte la [SL:ENTITY [IN:TIME_LAST dernière [SL:ENTITY échographie rénale]]]]] ?]
[IN:INTERPRETATION_VALUE quelle est la [SL:ENTITY [IN:TYPE_DATE_VALUE date et le résultat de la [SL:ENTITY [IN:TIME_LAST dernière [SL:ENTITY biopsie rénale]]]]] ?]
[IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TIME_ALL comment le [SL:ENTITY taux de neutrophiles] a-t-il évolué [SL:CONSTRAINT [IN:CONSTRAINT_RANGE entre [SL:DATE telle] et [SL:DATE telle date]]]]] ?]
[IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TIME_ALL comment le [SL:ENTITY taux de neutrophiles] a-t-il évolué [SL:CONSTRAINT [IN:CONSTRAINT_SINCE depuis l' [SL:DATE introduction du traitement xxx]]] ?]]]
[IN:INTERPRETATION_VALUE quelle est la [SL:ENTITY [IN:TYPE_DATE date de [SL:ENTITY sortie d' aplasie] [SL:CONSTRAINT [IN:CONSTRAINT_SINCE post [SL:DATE allo greffe de cellules souche hématopoïétiques]]]]]]
[IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TIME_ALL évolution de la [SL:ENTITY crp/procalcitonine] [SL:CONSTRAINT [IN:CONSTRAINT_SINCE depuis [SL:DATE 7 jours]]]]]]
[IN:INTERPRETATION_PRESENCE [SL:ENTITY [IN:TYPE_VALUE identification d' un [SL:ENTITY germe] [SL:CONSTRAINT [IN:CONSTRAINT_SINCE sur [SL:DATE 1 mois]]]]] ?]
[IN:INTERPRETATION_PRESENCE [SL:ENTITY [IN:TYPE_VALUE identification d' une [SL:ENTITY examen direct positif] [SL:CONSTRAINT [IN:CONSTRAINT_SINCE sur [SL:DATE 7 jours] ?]]]]]
[IN:INTERPRETATION_PRESENCE [SL:ENTITY [IN:TYPE_PRESENCE liste des prélèvements envoyés en [SL:ENTITY microbiologie] pour analyse]]]
[IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TIME_ALL suivi des [SL:ENTITY leucocytes] / [SL:ENTITY pnn] / [SL:ENTITY lymphocytes] / [SL:ENTITY monocytes] / [SL:ENTITY eosinophiles] [SL:CONSTRAINT [IN:CONSTRAINT_SINCE sur [SL:DATE 1 semaine]]]]]]
[IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TIME_ALL suivi de l' [SL:ENTITY hémoglobine] [SL:CONSTRAINT [IN:CONSTRAINT_SINCE sur [SL:DATE 3 jours]]]]]]
[IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TIME_ALL suivi [SL:ENTITY pcr ebv] [SL:CONSTRAINT [IN:CONSTRAINT_SINCE sur [SL:DATE 1 an]]]]]]
[IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TIME_ALL suivi [SL:ENTITY pcr cmv] [SL:CONSTRAINT [IN:CONSTRAINT_SINCE sur [SL:DATE 2 ans]]]]]]
[IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TIME_ALL suivi [SL:ENTITY clonalité t/b] [SL:CONSTRAINT [IN:CONSTRAINT_SINCE sur [SL:DATE 3 mois]]]]]]
[IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TIME_ALL suivi [SL:ENTITY bcr-abl] [SL:CONSTRAINT [IN:CONSTRAINT_SINCE sur [SL:DATE 15 jours]]]]]]
[IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TIME_ALL suivi des [SL:ENTITY lymphocytes] et [SL:ENTITY % de la population tumorale] [SL:CONSTRAINT [IN:CONSTRAINT_SINCE sur [SL:DATE 6 mois]]]]]]
[IN:INTERPRETATION_POSITIVITY [SL:ENTITY [IN:TIME_LAST dernière [SL:ENTITIY hémoculture]]] positive ?]
[IN:INTERPRETATION_POSITIVITY [SL:ENTITY [IN:TIME_ALL tableau des [SL:ENTITY [IN:TYPE_VALUE résultats positifs [SL:ENTITY bactériologiques]]]]]]
[IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TIME_ALL tableau des [SL:ENTITY [IN:TYPE_VALUE résultats de [SL:ENTITY cytométrie de flux] fait au [SL:SOURCE cedi]]]]]]
[IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TIME_ALL tableau des [SL:ENTITY [IN:TYPE_VALUE résultats de [SL:SOURCE génétique]]]]]]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat du [SL:ENTITY bilan hormonal] [SL:CONSTRAINT [IN:CONSTRAINT_DATE lors d' une [SL:DATE hypoglycémie]]] ?]]]
[IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TIME_ALL quel sont les [SL:ENTITY ammoniémies] [SL:CONSTRAINT [IN:CONSTRAINT_SINCE depuis [SL:DATE ce matin]]]]] ?]

Table C.1 : Questions 'vie réelle' *(continued)*

question
[IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TIME_ALL quelles sont les [SL:ENTITY [IN:TYPE_VALUE résultats du [SL:ENTITY cycle glycémie lactate] [SL:CONSTRAINT [IN:CONSTRAINT_SINCE depuis [SL:DATE hier]]]] ?]]]] [IN:INTERPRETATION_VALUE quelles sont les [SL:ENTITY [IN:TYPE_VALUE résultats de [SL:SOURCE génétique]]] de ce patient ?]
[IN:INTERPRETATION_VALUE quels sont les [SL:ENTITY [IN:TYPE_VALUE résultats de son [SL:ENTITY test au glucagon]] ?]] [IN:INTERPRETATION_VALUE quels sont ses [SL:ENTITY cpk]] ? [IN:INTERPRETATION_LOW sont [SL:ENTITY ils] en train de diminuer significativement ?] [IN:INTERPRETATION_POSITIVITY l' [SL:ENTITY hémoculture sur son ktc] est elle positive ?] [IN:INTERPRETATION_NORMAL comment était sa [SL:ENTITY [IN:TIME_LAST dernière [SL:ENTITY chromatographie des acides aminés plasmatiques]]] ?] [IN:INTERPRETATION_VALUE quel est son [SL:ENTITY trou anionique] ?]
[IN:INTERPRETATION_NONE [SL:ENTITY acuité visuelle]] [IN:INTERPRETATION_NONE [SL:ENTITY pression intraoculaire]] [IN:INTERPRETATION_NONE [SL:ENTITY épaisseur maculaire]] [IN:INTERPRETATION_VALUE quelle était la [SL:ENTITY [IN:TIME_LAST dernière [SL:ENTITY [IN:TYPE_VALUE valeur de l' [SL:ENTITY hémoglobine]]]]]] [IN:INTERPRETATION_VALUE quel était le [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY ecbu]]]]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY [IN:TYPE_VALUE compte rendu de [SL:ENTITY scanner cérébral]]]]]] [IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY compte rendu d' hospitalisation]]]] [IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat de la [SL:ENTITY [IN:TIME_LAST dernière [SL:ENTITY ponction lombaire]]]]] ?] [IN:INTERPRETATION_PRESENCE a t il déjà eu une [SL:ENTITY ponction lombaire] ?] [IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat des [SL:ENTITY pcr virales] dans le [SL:SOURCE lcr]]] ?]
[IN:INTERPRETATION_PRESENCE a t il un [SL:ENTITY syndrome inflammatoire] ?] [IN:INTERPRETATION_PRESENCE a t il eu une [SL:ENTITY recherche de virus] dans le [SL:SOURCE nez] ?] [IN:INTERPRETATION_PRESENCE a t il eu une [SL:ENTITY recherche de virus] dans les [SL:SOURCE selles]] [IN:INTERPRETATION_PRESENCE a t il eu une [SL:ENTITY recherche de mycoplasme]] [IN:INTERPRETATION_PRESENCE a t il eu un [SL:ENTITY [IN:TYPE_COMPLETENESS [SL:ENTITY bilan infectieux de maladie neuroinflammatoire] complet]]] ? [IN:INTERPRETATION_COMPLETENESS [SL:ENTITY [IN:TYPE_VALUE tous les resultats sont ils disponibles ?]]]
[IN:INTERPRETATION_PRESENCE a t il déjà eu une [SL:ENTITY recherche de glut1]] [IN:INTERPRETATION_PRESENCE a t il déjà eu des [SL:ENTITY chromatographies] du [SL:SOURCE sang] ? des [SL:SOURCE urines] ? du [SL:SOURCE lcr]] ? [IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TYPE_VALUE avec quels résultats]] ?] [IN:INTERPRETATION_VALUE quel est le résultat des [SL:ENTITY neurotransmetteurs] dans la [SL:SOURCE ponction lombaire]] [IN:INTERPRETATION_VALUE quel est le résultat de son [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY inr]]] ?] [IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TIME_ALL comment l' [SL:ENTITY inr] a évolué [SL:CONSTRAINT [IN:CONSTRAINT_SINCE depuis [SL:DATE un mois]]]]] ?]
[IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TIME_ALL comment l' [SL:ENTITY inr] a évolué [SL:CONSTRAINT [IN:CONSTRAINT_SINCE depuis [SL:DATE 6 mois]]]] ?]] [IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat de son [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY ntprobnp]]]]]] [IN:INTERPRETATION_VALUE comment le [SL:ENTITY ntprobnp] a évolué [SL:CONSTRAINT [IN:CONSTRAINT_SINCE depuis [SL:DATE 6 mois]]] ?] [IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat du [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY dosage de plaquettes]]]] ?]] [IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat du [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY antixa]]] ?]]]
[IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TIME_ALL comment l' [SL:ENTITY antixa] a évolué [SL:CONSTRAINT [IN:CONSTRAINT_SINCE depuis [SL:DATE 5 jours]]]]] ?] [IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat du [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY dosage de troponine]]]]] ?]

TABLE C.1 : Questions 'vie réelle' *(continued)*

question
[IN:INTERPRETATION_VALUE comment la [SL:ENTITY [IN:TIME_ALL [SL:ENTITY troponine] a évolué [SL:CONSTRAINT [IN:CONSTRAINT_SINCE depuis [SL:DATE 7 jours]]]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat du [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY dosage de transaminases]]]]] ?]
[IN:INTERPRETATION_VALUE comment les [SL:ENTITY [IN:TIME_ALL [SL:ENTITY transaminases] a évolué [SL:CONSTRAINT [IN:CONSTRAINT_SINCE depuis [SL:DATE 7 jours]]]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat de sa [SL:ENTITY [IN:TIME_LAST dernière [SL:ENTITY tsh]]]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat de sa [SL:ENTITY [IN:TIME_LAST dernière [SL:ENTITY hémoglobine]]]]] ?]
[IN:INTERPRETATION_VALUE comment la [SL:ENTITY [IN:TIME_ALL [SL:ENTITY crp] a évolué [SL:CONSTRAINT [IN:CONSTRAINT_SINCE depuis [SL:DATE 7 jours]]] ?]]]
[IN:INTERPRETATION_VALUE de quand date la [SL:ENTITY [IN:TYPE_DATE [SL:ENTITY [IN:TIME_LAST dernière [SL:ENTITY [IN:INTERPRETATION_POSITIVITY [SL:ENTITY [IN:TYPE_VALUE [SL:ENTITY hémoculture] positive]]]]]]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat du [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY dosage de tacrolimus]]]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat du [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY dosage de ciclosporine]]]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat du [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY dosage de cellcept]]]]] ?]
[IN:INTERPRETATION_PRESENCE prenez vous des [SL:ENTITY anticoagulants] ?]
[IN:INTERPRETATION_PRESENCE prenez vous des [SL:ENTITY antiaggrégants plaquettaires] ?]
[IN:INTERPRETATION_PRESENCE avez vous une [SL:ENTITY hypertension artérielle] ?]
[IN:INTERPRETATION_NORMAL la [SL:ENTITY [IN:TYPE_VALUE [SL:ENTITY fonction rénale]]] est elle normale ?]
[IN:INTERPRETATION_PRESENCE avez vous eu une ou plusieurs [SL:ENTITY colique néphérétique] ?]
[IN:INTERPRETATION_PRESENCE avez vous eu une [SL:ENTITY infection urinaire] ? [SL:ENTITY cystite] ou [SL:ENTITY pyélonéphrite] ?]
[IN:INTERPRETATION_VALUE quelle est votre [SL:ENTITY créatininémie] ?]
[IN:INTERPRETATION_VALUE connaissez vous votre [SL:ENTITY débit de filtration glomérulaire] ?]
[IN:INTERPRETATION_PRESENCE a t' il eu une [SL:ENTITY recherche de bmr] ou [SL:ENTITY bsle] ?]
[IN:INTERPRETATION_VALUE quelle est la [SL:ENTITY [IN:TYPE_DATE date de [SL:ENTITY recherche de bmr] ou [SL:ENTITY bsle]]] ?]
[IN:INTERPRETATION_VALUE quelle est l' [SL:ENTITY [IN:TIME_ALL évolution de la [SL:ENTITY bilirubine conjuguée]]] ?]
[IN:INTERPRETATION_VALUE quelle est l' [SL:ENTITY [IN:TIME_ALL évolution des [SL:ENTITY asat]]] ?]
[IN:INTERPRETATION_VALUE quelle est l' [SL:ENTITY [IN:TIME_ALL évolution des [SL:ENTITY alat]]] ?]
[IN:INTERPRETATION_VALUE quelle est l' [SL:ENTITY [IN:TIME_ALL évolution des [SL:ENTITY ggt]]] ?]
[IN:INTERPRETATION_VALUE quelle est la [SL:ENTITY [IN:TYPE_DATE date du [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY bilan biologique]]]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat de l' [SL:ENTITY électrophorèse des protéines]]] ?]
[IN:INTERPRETATION_HIGH a t' il plus de [SL:VALUE 50000] [SL:ENTITY [IN:TIME_LAST [SL:ENTITY plaquettes] au dernier bilan]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE nombre d' [SL:ENTITY éosinophiles]]] ?]
[IN:INTERPRETATION_PRESENCE y'a t' il une [SL:ENTITY hypoalbuminémie] ?]
[IN:INTERPRETATION_VALUE quelle est l' [SL:ENTITY [IN:TIME_ALL évolution de l' [SL:ENTITY antithrombine 3]]] ?]
[IN:INTERPRETATION_VALUE quelle est l' [SL:ENTITY [IN:TIME_ALL évolution du [SL:ENTITY tp]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE [SL:ENTITY résultat bactériologique] de la [SL:CONSTRAINT [IN:CONSTRAINT_DATE [SL:SOURCE biopsie] du [SL:DATE 4/02/17]]]]] ?]
[IN:INTERPRETATION_VALUE quelle est le [SL:ENTITY [IN:TYPE_VALUE résultat de la [SL:ENTITY [IN:TIME_LAST dernière [SL:ENTITY nfs]]]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_PRESENCE [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY bilan bio]]] réalisé]] ?]
[IN:INTERPRETATION_PRESENCE le patient a t il une [SL:ENTITY carte de groupe sanguin] ?]
[IN:INTERPRETATION_PRESENCE les [SL:ENTITY rai] ont elles été faites] et [IN:INTERPRETATION_UPTODATE sont [SL:ENTITY elles] à jour ?]

Table C.1 : Questions 'vie réelle' *(continued)*

question
[IN:INTERPRETATION_PRESENCE y a t il un [SL:ENTITY bilan d' hémostase] connu pour ce patient ?] [IN:INTERPRETATION_VALUE quels sont les [SL:ENTITY [IN:TYPE_VALUE résultats du [SL:ENTITY bilan d' hémostase]]] ?]
[IN:INTERPRETATION_VALUE quelle est la [SL:ENTITY [IN:TIME_ALL cinétique de la [SL:ENTITY numération plaquettaire] [SL:CONSTRAINT [IN:CONSTRAINT_SINCE sur les [SL:DATE 6 derniers mois]]]]] ?] [IN:INTERPRETATION_VALUE quels sont les [SL:ENTITY [IN:TYPE:VALUE résultats du [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY bilan]]]]]] [IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUEMAX chiffre maximal de [SL:ENTITY crp/pct]]]] [IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat de [SL:ENTITY tous les items de bio] disponibles dans [SL:SOURCE stare]]]] [IN:INTERPRETATION_PRESENCE le patient est il [SL:ENTITY [IN:TYPE_VALUE porteur de [SL:ENTITY bmr]]] ?]
[IN:INTERPRETATION_VALUE comment a [SL:ENTITY [IN:TIME_ALL évolué l' [SL:ENTITY antigénemie aspergillaire]]]] [IN:INTERPRETATION_VALUE comment a [SL:ENTITY [IN:TIME_ALL évolué le [SL:ENTITY taux de beta d glucane]]]] [IN:INTERPRETATION_VALUE de quand [SL:ENTITY [IN:TYPE_DATE date le [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY ecbu]]]]]] [IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résutlat du [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY prélèvemetn vaginal]]]]]] [IN:INTERPRETATION_NORMAL comment est la [SL:ENTITY clearance de la creatinine]]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY [IN:TYPE_VALUE résultat des [SL:ENTITY transaminases] ? des [SL:ENTITY gammagt] ? de la [SL:ENTITY bilirubine totale et conjuguée] ? de l' [SL:ENTITY albuminémie] ? des [SL:ENTITY gammaglobulines] ? du [SL:ENTITY tp] ? des [SL:ENTITY cofacteurs de la coagulation]]]]]] [IN:INTERPRETATION_VALUE quelle est l' [SL:ENTITY [IN:TIME_ALL évolution des [SL:ENTITY transaminases] ?cdes [SL:ENTITY gammagt] de la [SL:ENTITY bilirubine totale et conjuguée]]] ?] [IN:INTERPRETATION_POSITIVITY est-il [SL:ENTITY [IN:TYPE_VALUE immunisé vis à vis de l' [SL:ENTITY hépatite a]]] ?] [IN:INTERPRETATION_POSITIVITY est-il [SL:ENTITY [IN:TYPE_VALUE immunisée vis à vis de l' [SL:ENTITY hépatite b]]] ?] [IN:INTERPRETATION_VALUE quelle est le [SL:ENTITY [IN:TYPE_VALUE résultat du [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY dosage de l' alphafoetoprotéine sérique]]] ?]]]
[IN:INTERPRETATION_VALUE de quand [SL:ENTITY [IN:TYPE_DATE date le [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY dosage de l' alaphafoetoprotéine]]]]] ?] [IN:INTERPRETATION_VALUE de quand [SL:ENTITY [IN:TYPE_DATE date le [SL:ENTITY [IN:TIME_LAST dernier contrôle des [SL:ENTITY igg totaux anti hv] et des [SL:ENTITY ac anti hbs]]]]] ?] [IN:INTERPRETATION_PRESENCE y a t' il un résultat de [SL:ENTITY pcr ebv] ? [SL:ENTITY cmv] ? [SL:ENTITY hhv6] ?] [IN:INTERPRETATION_PRESENCE y a t' il une [SL:ENTITY hypergammaglobulinémie] ?] [IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat du [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY dosage de 25oh-vit d]]] ?]]]
[IN:INTERPRETATION_VALUE de quand [SL:ENTITY [IN:TYPE_DATE date le [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY dosage de 25oh-vit d]]]]] ?] [IN:INTERPRETATION_VALUE quelle est la [SL:ENTITY clearance de la créatinine] ?] [IN:INTERPRETATION_PRESENCE y a t' il des signes d' [SL:ENTITY insuffisance rénale] ?] [IN:INTERPRETATION_PRESENCE y a t' il des signes d' [SL:ENTITY infection] ?] [IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY taux d' albumine]]] ?]
[IN:INTERPRETATION_VALUE existe-t-il une [SL:ENTITY protéinurie significative] ?] [IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TIME_FIRST depuis [SL:ENTITY [IN:TYPE_DATE quand une [SL:ENTITY [IN:INTERPRETATION_PRESENCE [SL:ENTITY cytolyse] est-elle présente]] ?]]]]] [IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TIME_FIRST depuis [SL:ENTITY [IN:TYPE_DATE quand une [SL:ENTITY [IN:INTERPRETATION_PRESENCE [SL:ENTITY cholestase] est-elle présente]]] ?]]]] [IN:INTERPRETATION_PRESENCE existe-t-il une [SL:ENTITY carence martiale] ?] [IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY taux d' immunoglobulines]] ?]]

Table C.1 : Questions 'vie réelle' *(continued)*

question
[IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TIME_ALL evolution de l' [SL:ENTITY hemoglobine] du [SL:ENTITY taux de cd4] du [SL:ENTITY taux des gb] des [SL:ENTITY lymphocytes] [SL:CONSTRAINT [IN:CONSTRAINT_SINCE sur les [SL:DATE derniers mois]]]]] ?]
[IN:INTERPRETATION_PRESENCE y a t il une [SL:ENTITY documentation microbiologique] ?]
[IN:INTERPRETATION_PRESENCE est ce que les examens retrouvent un [SL:ENTITY germe] ?]
[IN:INTERPRETATION_VALUE est ce que les [SL:ENTITY [IN:TYPE_PRESENCE [SL:SOURCE cultures des examens microbiologiques] poussent à un [SL:ENTITY germe]]]]
[IN:INTERPRETATION_VALUE le [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY bilan mst]]] ?]
[IN:INTERPRETATION_VALUE la [SL:ENTITY [IN:TIME_LAST dernière [SL:ENTITY charge virale]]]]
[IN:INTERPRETATION_VALUE le [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY bilan mycologique]]]]
[IN:INTERPRETATION_VALUE le [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY dosage en igg]]]]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY [IN:TYPE_VALUE résultat des [SL:ENTITY anticorps anti transglutaminases]]]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY [IN:TYPE_VALUE résultat de la [SL:ENTITY calprotectine fécale]]]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat de la [SL:ENTITY recherche d' helicobacter pylori]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat d' [SL:ENTITY anapath] de la [SL:SOURCE [IN:TIME_LAST dernière [SL:SOURCE endoscopie]]]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat d' [SL:ENTITY anapath] de la [SL:SOURCE [IN:TIME_LAST dernière [SL:SOURCE fogd]]]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat d' [SL:ENTITY anapath] de la [SL:SOURCE [IN:TIME_LAST dernière [SL:SOURCE fibro]]]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat de [SL:ENTITY bacterio] de la [SL:SOURCE [IN:TIME_LAST dernière [SL:SOURCE endoscopie]]]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat de [SL:ENTITY bacterio] de la [SL:SOURCE [IN:TIME_LAST dernière [SL:SOURCE fogd]]]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat de [SL:ENTITY bacterio] de la [SL:SOURCE [IN:TIME_LAST dernière [SL:SOURCE fibro]]]]] ?]
[IN:INTERPRETATION_VALUE quels sont les [SL:ENTITY [IN:TYPE_VALUE résultats des [SL:ENTITY marqueurs fongiques]]] ?]
[IN:INTERPRETATION_VALUE quelle est l' [SL:ENTITY [IN:TIME_ALL évolution du [SL:ENTITY syndrome inflammatoire]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat de la [SL:ENTITY ponction lombaire]]] ?]
[IN:INTERPRETATION_VALUE quels sont les [SL:ENTITY [IN:TYPE_VALUE résultats du [SL:ENTITY lavage broncho alvéolaire]]] ?]
[IN:INTERPRETATION_VALUE quelle est [SL:ENTITY [IN:TIME_ALL lévolution de la [SL:ENTITY fonction rénale]]]]
[IN:INTERPRETATION_VALUE quelle est l' [SL:ENTITY [IN:TIME_ALL évolution du [SL:ENTITY bilan hépatique]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY bilan immuno virologique du vih]]] ?]
[IN:INTERPRETATION_VALUE le [SL:ENTITY [IN:TYPE_VALUE résultat du [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY génotypage de résistance]]]]] ?]
[IN:INTERPRETATION_VALUE quelle est la [SL:ENTITY ferritinémie] ?]
[IN:INTERPRETATION_VALUE quelle est la [SL:ENTITY calcémie] ?]
[IN:INTERPRETATION_VALUE a combien est la [SL:ENTITY crp] ?]
[IN:INTERPRETATION_VALUE quelle est son [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY traitement]]] ?]
[IN:INTERPRETATION_PRESENCE le patient reçoit-il du [SL:ENTITY cellcept] ?]
[IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TYPE_DATE quand est-il venu en [SL:ENTITY consultation] ?]]]
[IN:INTERPRETATION_PRESENCE le patient présente-t-il une [SL:ENTITY allergie] ?]

Table C.1 : Questions 'vie réelle' *(continued)*

question
[IN:INTERPRETATION_VALUE quel est son [SL:ENTITY diagnostic principal] ?]
[IN:INTERPRETATION_VALUE quelle est la [SL:ENTITY [IN:TYPE_DATE date du [SL:ENTITY diagnostic]]] ?]
[IN:INTERPRETATION_PRESENCE a-t-il un [SL:ENTITY abord veineux central] ?]
[IN:INTERPRETATION_VALUE quelle est la [SL:ENTITY [IN:TIME_ALL cinétique des [SL:ENTITY bhcg]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat de l' [SL:ENTITY amniocentèse]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY [IN:TYPE_VALUE résultat de la [SL:ENTITY cgh] / [SL:ENTITY caryotype]]] ?]
[IN:INTERPRETATION_VALUE quel est le [SL:ENTITY groupe sanguin] ?]
[IN:INTERPRETATION_POSITIVITY la patiente est elle [SL:ENTITY [IN:TYPE_VALUE immunisée pour la [SL:ENTITY toxoplasmose]]] ?]
[IN:INTERPRETATION_POSITIVITY la patiente est elle [SL:ENTITY [IN:TYPE_VALUE immunisée pour le [SL:ENTITY cmv]]] ?]
[IN:INTERPRETATION_PRESENCE mr x avait-il déjà une [SL:ENTITY thrombopénie] [SL:CONSTRAINT [IN:CONSTRAINT_RANGE lors de sa [SL:SOURCE [IN:TIME_LAST dernière [SL:SOURCE hospitalisation]]]]] ?]
[IN:INTERPRETATION_VALUE peux-tu me donner les [SL:ENTITY [IN:TYPE_VALUE résultats des [SL:ENTITY [IN:TIME_LAST [SL:ENTITY 4] dernières [SL:ENTITY créat']]]]] de mr x ?]
[IN:INTERPRETATION_HIGH quelle est la [SL:ENTITY [IN:TIME_ALL vitesse d' augmentation des [SL:ENTITY psa] de mr x [SL:CONSTRAINT [IN:CONSTRAINT_RANGE au cours des [SL:SOURCE [IN:TIME_LAST [SL:VALUE 4] dernières [SL:SOURCE hospitalisations]]]]]]] ?]
[IN:INTERPRETATION_LOW la [SL:ENTITY [IN:TIME_ALL [SL:ENTITY crp] de mme y à t' elle diminuée]] ?]
[IN:INTERPRETATION_PRESENCE le [SL:CONSTRAINT [IN:CONSTRAINT_DATE [SL:SOURCE [IN:TIME_LAST dernier [SL:SOURCE bilan]]]]] de mme y est-il compatible avec une [SL:ENTITY anémie hémolitique auto-immune] ?]
[IN:INTERPRETATION_PRESENCE des [SL:ENTITY carences] ont-elles été mises en évidence dans les [SL:SOURCE [IN:TIME_LAST derniers [SL:SOURCE bilan]]] de mme y ?]
[IN:INTERPRETATION_VALUE peux-tu me donner la [SL:ENTITY [IN:TYPE_DATE_VALUE date et la valeur du [SL:ENTITY [IN:TIME_LAST dernier [SL:ENTITY [IN:INTERPRETATION_NORMAL [SL:ENTITY taux de plaquette] normal]]]]]] ?]
[IN:INTERPRETATION_VALUE peux-tu me donner la [SL:ENTITY [IN:TYPE_DATE_VALUE date et la valeur de la [SL:ENTITY [IN:TIME_FIRST première [SL:ENTITY crp]]]]] de mme y ?]
[IN:INTERPRETATION_PRESENCE a-t-on doser des [SL:ENTITY marqueurs tumoraux] pour mme y ?]
[IN:INTERPRETATION_PRESENCE le [SL:SOURCE [IN:TIME_LAST dernier [SL:SOURCE bilan]]] de mr x est-il compatible avec une [SL:ENTITY déshydratation extracellulaire] ?]
[IN:INTERPRETATION_VALUE quelle est la [SL:ENTITY [IN:TYPE_VALUE valeur de la [SL:ENTITY natrémie corrigée]]] de mr x ?]
[IN:INTERPRETATION_VALUE quelle est la [SL:ENTITY [IN:TYPE_VALUEMAX valeur du pic de [SL:ENTITY glycémie] de mme y [SL:CONSTRAINT [IN:CONSTRAINT_RANGE au cours de son [SL:DATE hospitalisation]]]] ?]]
[IN:INTERPRETATION_VALUE quelle est la [SL:ENTITY [IN:TYPE_VALUEMEAN moyenne des [SL:ENTITY hémoglobines glyquées] de mr x [SL:CONSTRAINT [IN:CONSTRAINT_SINCE pour [SL:DATE cette année]]]]] ?]
[IN:INTERPRETATION_VALUE quelle est la [SL:ENTITY [IN:TYPE_VALUEMEAN valeur moyenne de la [SL:ENTITY glycémie] [SL:CONSTRAINT [IN:CONSTRAINT_RANGE pour [SL:DATE chacune des 4 dernières hospitalisations]]]] de mr x] ?]
[IN:INTERPRETATION_VALUE les [SL:ENTITY [IN:TIME_ALL [SL:ENTITY marqueurs tumoraux] de mr x sont-ils stables [SL:CONSTRAINT [IN:CONSTRAINT_SINCE au cours des [SL:DATE 6 derniers mois]]]]] ?]
[IN:INTERPRETATION_VALUE [SL:ENTITY [IN:TIME_ALL l' [SL:ENTITY insuffisance rénale] de mr x est-elle stable [SL:CONSTRAINT [IN:CONSTRAINT_SINCE [SL:DATE cette année]]]]] ?]
[IN:INTERPRETATION_VALUE peux-tu m'afficher le [SL:ENTITY [IN:TIME_ALL graph' de l' [SL:ENTITY urée] et de la [SL:ENTITY creat'] [SL:CONSTRAINT [IN:CONSTRAINT_RANGE au cours de [SL:DATE cette année]]]] pour mme y] ?]
[IN:INTERPRETATION_VALUE peux tu me donner [SL:ENTITY [IN:TYPE_VALUEMEAN la valeur moyenne du [SL:ENTITY vgm] [SL:CONSTRAINT [IN:CONSTRAINT_RANGE au cours de la [SL:DATE pénultième hospitalisation]]]] pour mme y] ?]

TABLE C.1 : Questions 'vie réelle' *(continued)*

question
[IN:INTERPRETATION_VALUE peux tu me donner le [SL:ENTITY taux d' ige] [SL:CONSTRAINT [IN:CONSTRAINT_RANGE au cours de l' [SL:DATE avant avant dernière hospitalisation]]] pour mme y ?]
[IN:INTERPRETATION_VALUE quelle est la [SL:ENTITY [IN:TYPE_DATE date de la [SL:ENTITY [IN:TIME_LAST dernière [SL:ENTITY intervention chirurgicale]]]]] ?] [IN:INTERPRETATION_VALUE de quand [SL:ENTITY [IN:TYPE_DATE date la [SL:ENTITY [IN:TIME_LAST dernière [SL:ENTITY échographie abdominale]]]]] ?]

References

00 1. Hill RG, Sears LM, Melanson SW. 4000 Clicks : A productivity analysis of electronic medical records in a community hospital ED. The American Journal of Emergency Medicine [Internet]. 2013 Nov [cited 2020 Oct 14] ;31(11):1591–4. Available from : http://www.sciencedirect.com/science/article/pii/S0735675713004051

2. Sittig DF, Murphy DR, Smith MW, Russo E, Wright A, Singh H. Graphical display of diagnostic test results in electronic health records : A comparison of 8 systems. Journal of the American Medical Informatics Association : JAMIA. 2015 Jul ;22(4):900–4.

3. Charon R. At the Membranes of Care : Stories in Narrative Medicine. Academic Medicine [Internet]. 2012 Mar [cited 2020 Oct 14] ;87(3):342–7. Available from : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3292868/

4. Poissant L, Pereira J, Tamblyn R, Kawasumi Y. The impact of electronic health records on time efficiency of physicians and nurses : A systematic review. Journal of the American Medical Informatics Association : JAMIA. 2005 Sep-Oct ;12(5):505–16.

5. Gardner RL, Cooper E, Haskell J, Harris DA, Poplau S, Kroth PJ, et al. Physician stress and burnout : The impact of health information technology. Journal of the American Medical Informatics Association [Internet]. 2019 Feb [cited 2020 Oct 14] ;26(2):106–14. Available from : https://academic.oup.com/jamia/article/26/2/106/5230918

6. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes : A perspective on the tension between structure and flexible documentation. Journal of the American Medical Informatics Association : JAMIA. 2011 Mar-Apr ;18(2):181–6.

7. Spyns P. Natural language processing in medicine : An overview [Internet]. Vol. 35, Methods of information in medicine. Methods Inf Med ; 1996 [cited 2020 Oct 14]. Available from : https://pubmed.ncbi.nlm.nih.gov/9019092/

8. Meystre Sm, Gk S, Kc K-S, Jf H. Extracting information from textual documents in the electronic health record : A review of recent research [Internet]. Yearbook of medical informatics. Yearb Med Inform ; 2008 [cited 2020 Oct 14]. Available from : https://pubmed.ncbi.nlm.nih.gov/18660887/

9. Yim W, Yetisgen M, Harris WP, Kwan SW. Natural Language Processing in Oncology : A Review. JAMA Oncology [Internet]. 2016 Jun [cited 2020 Oct 14] ;2(6):797–804. Available from : https://jamanetwork.com/journals/jamaoncology/fullarticle/2517402

10. Lafferty J, McCallum A, Pereira F. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In : Proceedings of the international conference on machine learning (ICML). 2001. p. 282–9.

11. Sutton C, McCallum A. An Introduction to Conditional Random Fields. arXiv:10114088 [stat] [Internet]. 2010 Nov [cited 2020 Oct 15] ; Available from : http://arxiv.org/abs/1011.4088

12. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics [Internet]. 1943 Dec [cited 2020 Oct 14] ;5(4):115–33. Available from : https://doi.org/10.1007/BF02478259

13. Minsky M, Papert S. Perceptrons, An Introduction to Computational Geometry [Internet]. MIT Press. The MIT Press ; 1969 [cited 2020 Oct 14]. Available from : https://mitpress.mit.edu/books/perceptrons

14. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. In : Proceedings of the 27th International Conference on International Conference on Machine Learning. Madison, WI, USA : Omnipress ; 2010. p. 807–14. (ICML'10).

15. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature [Internet]. 1986 Oct [cited 2020 Oct 14] ;323(6088):533–6. Available from : https://www.nature.com/articles/323533a0

16. wikipedia. Rétropropagation du gradient. Wikipédia [Internet]. 2020 Oct [cited 2020 Oct 14] ; Available from : https://fr.wikipedia.org/w/index.php?title=R/%C3/%A9tropropagation_du_gradient/&oldid=175560157

17. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998 Nov ;86(11):2278–324.

18. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation [Internet]. 1997 Nov [cited 2018 Nov 17] ;9(8):1735–80. Available from : https://doi.org/10.1162/neco.1997.9.8.1735

19. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv:14123555 [cs] [Internet]. 2014 Dec [cited 2020 Oct 15] ; Available from : http://arxiv.org/abs/1412.3555

20. Schuster M, Paliwal K. Bidirectional recurrent neural networks. Signal Processing, IEEE Transactions on. 1997 Dec ;45:2673–81.

21. Sundermeyer M, Alkhouli T, Wuebker J, Ney H. Translation Modeling with Bidirectional Recurrent Neural Networks. In : Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) [Internet]. Doha, Qatar : Association for Computational Linguistics ; 2014 [cited 2020 Oct 15]. p. 14–25. Available from : https://www.aclweb.org/anthology/D14-1003

22. Kiperwasser E, Goldberg Y. Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations. Transactions of the Association for Computational Linguistics [Internet]. 2016 [cited 2020 Oct 15] ;4:313–27. Available from : https://www.aclweb.org/anthology/Q16-1023

23. Dernoncourt F, Lee JY, Szolovits P. NeuroNER : An easy-to-use program for named-entity recognition based on neural networks. arXiv:170505487 [cs, stat] [Internet]. 2017 May [cited 2020 Oct 15] ; Available from : http://arxiv.org/abs/1705.05487

24. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:14090473 [cs, stat] [Internet]. 2016 May [cited 2020 Oct 15] ; Available from : http://arxiv.org/abs/1409.0473

25. Graves A, Wayne G, Danihelka I. Neural Turing Machines. arXiv:14105401 [cs] [Internet]. 2014 Dec [cited 2020 Oct 15] ; Available from : http://arxiv.org/abs/1410.5401

26. Luong M-T, Pham H, Manning CD. Effective Approaches to Attention-based Neural Machine Translation. arXiv:150804025 [cs] [Internet]. 2015 Aug [cited 2017 Oct 16] ; Available from : http://arxiv.org/abs/1508.04025

27. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. arXiv:170603762 [cs] [Internet]. 2017 Jun [cited 2018 Oct 31] ; Available from : http://arxiv.org/abs/1706.03762

28. Xu K, Lei J, Kiros R, Cho K, Courville A, Salakhutdinov R, et al. Show, Attend and Tell : Neural Image CaptionGeneration with Visual Attention. In : Proceedings of the32ndInternational Conference on MachineLearning. Lille, France ; 2015. p. 10.

29. Cheng J, Dong L, Lapata M. Long Short-Term Memory-Networks for Machine Reading. arXiv:160106733 [cs] [Internet]. 2016 Sep [cited 2020 Oct 15] ; Available from : http://arxiv.org/abs/1601.06733

30. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and their Compositionality. In : Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. Advances in Neural Information Processing Systems 26 [Internet]. Curran Associates, Inc. ; 2013 [cited 2017 Mar 31]. p. 3111–9. Available from : http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

31. Pennington J, Socher R, Manning CD. GloVe : Global Vectors for Word Representation. In : Empirical Methods in Natural Language Processing (EMNLP) [Internet]. 2014. p. 1532–43. Available from : http://www.aclweb.org/anthology/D14-1162

32. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. arXiv preprint arXiv:160704606 [Internet]. 2016 [cited 2017 Feb 14] ; Available from : https://arxiv.org/abs/1607.04606

33. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep Contextualized Word Representations. In : Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers) [Internet]. New Orleans, Louisiana : Association for Computational Linguistics ; 2018 [cited 2018 Oct 5]. p. 2227–37. Available from : http://aclweb.org/anthology/N18-1202

34. Devlin J, Chang M-W, Lee K, Toutanova K. BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:181004805 [cs] [Internet]. 2018 Oct [cited 2019 Sep 13] ; Available from : http://arxiv.org/abs/1810.04805

35. Clark K, Luong M-T, Le QV. ELECTRA : PRE-TRAINING TEXT ENCODERS AS DISCRIMINATORS RATHER THAN GENERATORS. 2020 ;18.

36. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT : A Lite BERT for Self-supervised Learning of Language Representations. arXiv:190911942 [cs] [Internet]. 2020 Feb [cited 2020 Oct 15] ; Available from : http://arxiv.org/abs/1909.11942

37. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa : A Robustly Optimized BERT Pretraining Approach. arXiv:190711692 [cs] [Internet]. 2019 Jul [cited 2020 Oct 3] ; Available from : http://arxiv.org/abs/1907.11692

38. Le H, Vial L, Frej J, Segonne V, Coavoux M, Lecouteux B, et al. FlauBERT : Unsupervised Language Model Pre-training for French. arXiv:191205372 [cs] [Internet]. 2020 Mar [cited 2020 Oct 15] ; Available from : http://arxiv.org/abs/1912.05372

39. Martin L, Muller B, Ortiz Suárez PJ, Dupont Y, Romary L, de la Clergerie É, et al. CamemBERT : A Tasty French Language Model. In : Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics [Internet]. Online : Association for Computational Linguistics ; 2020 [cited 2020 Oct 3]. p. 7203–19. Available from : https://www.aclweb.org/anthology/2020.acl-main.645

40. Weizenbaum J. ELIZA - a computer program for the study of natural language communication between man and machine. Communications of the ACM [Internet]. 1966 Jan [cited 2020 Oct 15] ;9(1):36–45. Available from : https://doi.org/10.1145/365153.365168

41. Watson A, Bickmore T, Cange A, Kulshreshtha A, Kvedar J. An Internet-Based Virtual Coach to Promote Physical Activity Adherence in Overweight Adults : Randomized Controlled Trial. Journal of Medical Internet Research [Internet]. 2012 [cited 2020 Oct 15] ;14(1):e1. Available from : https://www.jmir.org/2012/1/e1/

42. Bickmore TW, Schulman D, Sidner C. Automated interventions for multiple health behaviors using conversational agents. Patient Education and Counseling [Internet]. 2013 Aug [cited 2020 Oct 15] ;92(2):142–8. Available from : http://www.sciencedirect.com/science/article/pii/S0738399113002115

43. Bickmore TW, Silliman RA, Nelson K, Cheng DM, Winter M, Henault L, et al. A Randomized Controlled Trial of an Automated Exercise Coach for Older Adults. Journal of the American Geriatrics Society [Internet]. 2013 [cited 2020 Oct 15] ;61(10):1676–83. Available from : https://onlinelibrary.wiley.com/doi/abs/10.1111/jgs.12449

44. Radziwill N, Benton M. Evaluating Quality of Chatbots and Intelligent Conversational Agents. 2017 Apr ;

45. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare : A systematic review. Journal of the American Medical Informatics Association. 2018 ;

46. D'Alfonso S, Santesteban-Echarri O, Rice S, Wadley G, Lederman R, Miles C, et al. Artificial Intelligence-Assisted Online Social Therapy for Youth Mental Health. Frontiers in Psychology [Internet]. 2017 [cited 2020 Oct 11] ;8. Available from : https://www.frontiersin.org/articles/10.3389/fpsyg.2017.00796/full

47. Bickmore TW, Schulman D, Sidner CL. A reusable framework for health counseling dialogue systems based on a behavioral medicine ontology. Journal of Biomedical Informatics [Internet]. 2011 Apr [cited 2020 Oct 11] ;44(2):183–97. Available from : http://www.sciencedirect.com/science/article/pii/S1532046411000025

48. Ring L, Bickmore T, Pedrelli P. An Affectively Aware Virtual Therapist for Depression Counseling. :4.

49. Fitzpatrick KK, Darcy A, Vierhile M. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot) : A Randomized Controlled Trial. JMIR Mental Health [Internet]. 2017 [cited 2020 Oct 11] ;4(2):e19. Available from : https://mental.jmir.org/2017/2/e19/

50. Miner AS, Milstein A, Schueller S, Hegde R, Mangurian C, Linos E. Smartphone-Based Conversational Agents and Responses to Questions About Mental Health, Interpersonal Violence, and Physical Health. JAMA Internal Medicine [Internet]. 2016 May [cited 2020 Oct 11] ;176(5):619. Available from : http://archinte.jamanetwork.com/article.aspx?doi=10.1001/jamainternmed.2016.0400

51. Ireland D, Atay C, Liddle J(Jacqueline), Bradford D, Lee H, Rushin O, et al. Hello Harlie : Enabling speech monitoring through chat-bot conversations. Studies in Health Technology and Informatics [Internet]. 2016 [cited 2020 Oct 11] ;227:55–60. Available from : `https://eprints.qut.edu.au/128687/`

52. Rhee H, Allen J, Mammen J, Swift M. Mobile phone-based asthma self-management aid for adolescents (mASMAA) : A feasibility study. Patient preference and adherence [Internet]. 2014 Jan [cited 2020 Oct 11] ;8:63–72. Available from : `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3891581/`

53. Tanaka H, Negoro H, Iwasaka H, Nakamura S. Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. PLOS ONE [Internet]. 2017 Aug [cited 2020 Oct 11] ;12(8):e0182151. Available from : `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0182151`

54. Nikitina S, Callaioli S, Baez M. Smart Conversational Agents for Reminiscence. In : 2018 IEEE/ACM 1st International Workshop on Software Engineering for Cognitive Services (SE4COG). 2018. p. 52–7.

55. Lisetti C, Yasavur U. Building an On-Demand Avatar-Based Health Intervention for Behavior Change. :6.

56. Kenny P, Parsons T. Embodied conversational virtual patients. Conversational Agents and Natural Language Interaction : Techniques and Effective Practices Information Science Reference. 2011 ;254–81.

57. Rizzo A, Kenny P, Parsons TD. Intelligent Virtual Patients for Training Clinical Skills. JVRB - Journal of Virtual Reality and Broadcasting [Internet]. 2011 Feb ;8(2011)(3). Available from : `http://www.jvrb.org/past-issues/8.2011/2902`

58. Chuah JH, Robb A, White C, Wendling A, Lampotang S, Kopper R, et al. Exploring Agent Physicality and Social Presence for Medical Team Training. Presence : Teleoperators and Virtual Environments [Internet]. 2013 Aug [cited 2020 Oct 11] ;22(2):141–70. Available from : `https://doi.org/10.1162/PRES_a_00145`

59. Campillos-Llanos L, Bouamor D, Bilinski É, Ligozat A-L, Zweigenbaum P, Rosset S. Description of the PatientGenesys dialogue system. In : Proc Of SIGDIAL [Internet]. Prague, Czech Republic : ACL ; 2015. p. 438–40. Available from : `http://aclweb.org/anthology/W15-4660`

60. Talbot TB, Kalisch N, Christoffersen K, Lucas G, Forbell E. Natural language understanding performance and use considerations in virtual medical encounters. Medicine Meets Virtual Reality 22 : NextMed/MMVR22. 2016 ;220:407–13.

61. Tanaka H, Adachi H, Ukita N, Ikeda M, Kazui H, Kudo T, et al. Detecting Dementia Through Interactive Computer Avatars. IEEE Journal of Translational Engineering in Health and Medicine. 2017 ;5:1–1.

62. Wargnier P, Carletti G, Laurent-Corniquet Y, Benveniste S, Jouvelot P, Rigaud A-S. Field evaluation with cognitively-impaired older adults of attention management in the Embodied Conversational Agent Louise. In : 2016 IEEE International Conference on Serious Games and Applications for Health (SeGAH). 2016. p. 1–8.

63. Kowatsch T, Nißen M, Shih C-HI, Rüegger D, Volland D, Filler A, et al. Text-based Healthcare Chatbots Supporting Patient and Health Professional Teams : Preliminary Results of a Randomized Controlled Trial on Childhood Obesity. In : Persuasive Embodied Agents for Behavior Change (PEACH2017) Workshop, co-located with the 17th International Conference on Intelligent Virtual Agents (IVA 2017) [Internet]. Stockholm, Sweden ; 2017 [cited 2020 Oct 11]. Available from : `https://www.researchgate.net/publication/320161507`

64. Sebastian J, Richards D. Changing stigmatizing attitudes to mental health via education and contact with embodied conversational agents. Computers in Human Behavior [Internet]. 2017 Aug [cited 2020 Oct 11] ;73:479–88. Available from : `http://www.sciencedirect.com/science/article/pii/S0747563217302467`

65. Tielman ML, Neerincx MA, van Meggelen M, Franken I, Brinkman W-P. How should a virtual agent present psychoeducation ? Influence of verbal and textual presentation on adherence. Technology and Health Care [Internet]. 2017 Jan [cited 2020 Oct 11] ;25(6):1081–96. Available from : `https://content.iospress.com/articles/technology-and-health-care/thc170899`

66. Zhang Z, Bickmore T, Mainello K, Mueller M, Foley M, Jenkins L, et al. Maintaining Continuity in Longitudinal, Multi-method Health Interventions Using Virtual Agents : The Case of Breastfeeding Promotion. In : Bickmore T, Marsella S, Sidner C, editors. Intelligent Virtual Agents. Cham : Springer International Publishing ; 2014. p. 504–13. (Lecture Notes in Computer Science).

67. Alesanco Á, Sancho J, Gilaberte Y, Abarca E, García J. Bots in messaging platforms, a new paradigm in healthcare delivery : Application to custom prescription in dermatology. In : Eskola H, Väisänen O, Viik J, Hyttinen J, editors. EMBEC & NBC 2017. Singapore : Springer ; 2018. p. 185–8. (IFMBE Proceedings).

68. King AC, Bickmore TW, Campero MI, Pruitt LA, Yin JL. Employing Virtual Advisors in Preventive Care for Underserved Communities : Results From the COMPASS Study. Journal of Health Communication [Internet]. 2013 Dec [cited 2020 Oct 11] ;18(12):1449–64. Available from : `https://doi.org/10.1080/10810730.2013.798374`

69. Bresó A, Martínez-Miranda J, Botella C, Baños RM, García-Gómez JM. Usability and acceptability assessment of an empathic virtual agent to prevent major depression. Expert Systems [Internet]. 2016 [cited 2020 Oct 11] ;33(4):297–312. Available from : `https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.12151`

70. Ni L, Lu C, Liu N, Liu J. MANDY : Towards a Smart Primary Care Chatbot Application. In : Chen J, Theeramunkong T, Supnithi T, Tang X, editors. Knowledge and Systems Sciences. Singapore : Springer ; 2017. p. 38–52. (Communications in Computer and Information Science).

71. Nugues P, ElGuedj P-O, Cazenave F, de Ferrière B. Issues in the design of a voice man machine dialogue system generating written medical reports. In : Engineering in medicine and biology society, 1992 14th annual international conference of the IEEE. IEEE ; 1992. p. 842–4.

72. Schaik V, Ds S, WA C. Evaluating a spoken dialogue system for recording clinical observations during an endoscopic examination. Medical informatics and the Internet in medicine. 2003 ;

73. Sonntag D, Möller M. Unifying semantic annotation and querying in biomedical image repositories. In : Proceedings of international conference on knowledge management and information sharing (KMIS). 2009.

74. Beveridge M, Fox J. Automatic generation of spoken dialogue from medical plans and ontologies. Journal of Biomedical Informatics. 2006 ;39(5):482–99.

75. Sonntag D, Schulz C. A multimodal multi-device discourse and dialogue infrastructure for collaborative decision-making in medicine. In : Mariani J, Rosset S, Garnier-Rizet M, Devillers L, editors. Natural interaction with robots, knowbots and smartphones : Putting spoken dialog systems into practice. New York, NY : Springer New York ; 2014. p. 37–47.

76. Philip P, Micoulaud-Franchi J-A, Sagaspe P, De Sevin E, Olive J, Bioulac S, et al. Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders. Scientific Reports. 2017 ;7:42656.

77. Lucas GM, Rizzo A, Gratch J, Scherer S, Stratou G, Boberg J, et al. Reporting Mental Health Symptoms : Breaking Down Barriers to Care with Virtual Human Interviewers. Frontiers in Robotics and AI [Internet]. 2017 [cited 2020 Oct 11] ;4. Available from : `https://www.frontiersin.org/articles/10.3389/frobt.2017.00051/full`

78. Philip P, Bioulac S, Sauteraud A, Chaufton C, Olive J. Could a Virtual Human Be Used to Explore Excessive Daytime Sleepiness in Patients ? Presence : Teleoperators and Virtual Environments [Internet]. 2014 Nov [cited 2020 Oct 11] ;23(4):369–76. Available from : `https://doi.org/10.1162/PRES_a_00197`

79. Jokinen K, McTear M. Spoken dialogue systems. Morgan & Claypool Publishers ; 2009. (Synthesis lectures on human language technologies ; vol. 2).

80. Chomsky N. Aspects of the Theory of Syntax. MIT Press ; 1969.

81. Price PJ. Evaluation of spoken language systems : The ATIS domain. In : Proceedings of the workshop on Speech and Natural Language [Internet]. USA : Association for Computational Linguistics ; 1990 [cited 2020 Oct 16]. p. 91–5. (HLT '90). Available from : `https://doi.org/10.3115/116580.116612`

82. Bonneau-Maynard H, Rosset S, Ayache C, Kuhn A, Mostefa D. Semantic annotation of the French Media Dialog Corpus. 2005.

83. Pieraccini R, Tzoukermann E, Gorelov Z, Gauvain J-L, Levin E, Lee C-H, et al. A speech understanding system based on statistical representation of semantics. In : Proceedings of the 1992 IEEE international conference on Acoustics, speech and signal processing - Volume 1. USA : IEEE Computer Society ; 1992. p. 193–6. (ICASSP'92).

84. Kuhn R, De Mori R. The application of semantic classification trees to natural language understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1995 May ;17(5):449–60.

85. Seneff S. TINA : A Natural Language System for Spoken Language Applications. Computational Linguistics [Internet]. 1992 [cited 2020 Oct 16] ;18(1):61–86. Available from : `https://www.aclweb.org/anthology/J92-1004`

86. Ward W, Issar S. Recent Improvements in the CMU Spoken Language Understanding System. In : Human Language Technology : Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994 [Internet]. 1994 [cited 2020 Oct 16]. Available from : `https://www.aclweb.org/anthology/H94-1039`

87. Raymond C, Riccardi G. Generative and discriminative algorithms for spoken language understanding. 2007.

88. Tur G, Hakkani-Tur D, Heck L. What is left to be understood in ATIS ? 2011.

89. Tur G, Mori RD. Spoken Language Understanding : Systems for Extracting Semantic Information from Speech. John Wiley & Sons ; 2011.

90. Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. Neural Computation [Internet]. 2006 Jul [cited 2020 Oct 16] ;18(7):1527–54. Available from : https://doi.org/10.1162/neco.2006.18.7.1527

91. Collobert R, Weston J. A unified architecture for natural language processing : Deep neural networks with multitask learning. In : Proceedings of the 25th international conference on Machine learning [Internet]. New York, NY, USA : Association for Computing Machinery ; 2008 [cited 2020 Oct 16]. p. 160–7. (ICML '08). Available from : https://doi.org/10.1145/1390156.1390177

92. Kim Y. Convolutional Neural Networks for Sentence Classification. In : Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) [Internet]. Doha, Qatar : Association for Computational Linguistics ; 2014 [cited 2020 Oct 16]. p. 1746–51. Available from : https://www.aclweb.org/anthology/D14-1181

93. Kalchbrenner N, Grefenstette E, Blunsom P. A Convolutional Neural Network for Modelling Sentences. In : Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers) [Internet]. Baltimore, Maryland : Association for Computational Linguistics ; 2014 [cited 2020 Oct 16]. p. 655–65. Available from : https://www.aclweb.org/anthology/P14-1062

94. Ravuri SV, Stolcke A. Recurrent neural network and LSTM models for lexical utterance classification. In : INTERSPEECH [Internet]. 2015 [cited 2017 Sep 1]. p. 135–9. Available from : http://ai2-s2-pdfs.s3.amazonaws.com/98d7/071f5dbfdc413e9faf06b1db91531e2c2c61.pdf

95. Hakkani-Tur D, Tur G, Asli C, Chen Y-N, Gao J, Deng li, et al. Multi-Domain Joint Semantic Frame Parsing using Bi-directional RNN-LSTM. 2016.

96. Lee JY, Dernoncourt F. Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. arXiv:160303827 [cs, stat] [Internet]. 2016 Mar [cited 2017 Sep 28] ; Available from : http://arxiv.org/abs/1603.03827

97. Zhao Z, Wu Y. Attention-Based Convolutional Neural Networks for Sentence Classification. In : Interspeech 2016 [Internet]. 2016 [cited 2020 Oct 16]. p. 705–9. Available from : http://www.isca-speech.org/archive/Interspeech_2016/abstracts/0354.html

98. Yang Z, Yang D, Dyer C, He X, Smola AJ, Hovy EH. Hierarchical Attention Networks for Document Classification. In : HLT-NAACL [Internet]. 2016 [cited 2017 Sep 29]. p. 1480–9. Available from : http://www.aclweb.org/anthology/N16-1174

99. Vu NT. Sequential Convolutional Neural Networks for Slot Filling in Spoken Language Understanding. In : Interspeech 2016 [Internet]. 2016 [cited 2020 Oct 16]. p. 3250–4. Available from : http://www.isca-speech.org/archive/Interspeech_2016/abstracts/0395.html

100. Mesnil G, He X, Deng L, Bengio Y. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In : INTERSPEECH. 2013. p. 3771–5.

101. Yao K, Peng B, Zhang Y, Yu D, Zweig G, Shi Y. Spoken language understanding using long short-term memory neural networks. In : 2014 IEEE Spoken Language Technology Workshop (SLT). 2014. p. 189–94.

102. Peng B, Yao K. Recurrent Neural Networks with External Memory for Language Understanding. arXiv:150600195 [cs] [Internet]. 2015 May [cited 2020 Oct 16] ; Available from : http://arxiv.org/abs/1506.00195

103. Kurata G, Xiang B, Zhou B, Yu M. Leveraging Sentence-level Information with Encoder LSTM for Semantic Slot Filling. In : Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing [Internet]. Austin, Texas : Association for Computational Linguistics ; 2016 [cited 2020 Oct 16]. p. 2077–83. Available from : https://www.aclweb.org/anthology/D16-1223

104. Zhao L, Feng Z. Improving Slot Filling in Spoken Language Understanding with Joint Pointer and Attention. In : Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers) [Internet]. Melbourne, Australia : Association for Computational Linguistics ; 2018 [cited 2020 Oct 16]. p. 426–31. Available from : https://www.aclweb.org/anthology/P18-2068

105. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information : A systematic review. Journal of biomedical informatics [Internet]. 2017 Sep [cited 2020 Sep 1] ;73:14–29. Available from : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6864736/

106. Spasic I, Nenadic G. Clinical Text Data in Machine Learning : Systematic Review. JMIR Medical Informatics [Internet]. 2020 [cited 2020 Oct 14] ;8(3):e17984. Available from : https://medinform.jmir.org/2020/3/e17984/

107. Fu S, Chen D, He H, Liu S, Moon S, Peterson KJ, et al. Clinical concept extraction : A methodology review. Journal of Biomedical Informatics [Internet]. 2020 Sep [cited 2020 Oct 16] ;109:103526. Available from : http://www.sciencedirect.com/science/article/pii/S1532046420301544

108. Hahn U, Oleynik M. Medical Information Extraction in the Age of Deep Learning. Yearbook of Medical Informatics [Internet]. 2020 Aug [cited 2020 Sep 1] ;29(1):208–20. Available from : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7442512/

109. DREES. Les depenses de sante en 2018 [Internet]. DREES ; 2019 [cited 2020 Sep 1]. Available from : https://drees.solidarites-sante.gouv.fr/IMG/pdf/cns2019.pdf

110. Pirmohamed M, James S, Meakin S, Green C, Scott AK, Walley TJ, et al. Adverse drug reactions as cause of admission to hospital : Prospective analysis of 18 820 patients. BMJ (Clinical research ed). 2004 Jul ;329(7456):15–9.

111. Olivier P, Boulbés O, Tubery M, Lauque D, Montastruc J-L, Lapeyre-Mestre M. Assessing the feasibility of using an adverse drug reaction preventability scale in clinical practice : A study in a French emergency department. Drug Safety. 2002 ;25(14):1035–44.

112. Zhou L, Mahoney LM, Shakurova A, Goss F, Chang FY, Bates DW, et al. How many medication orders are entered through free-text in EHRs ?–a study on hypoglycemic agents. AMIA Annual Symposium proceedings AMIA Symposium. 2012 ;2012:1079–88.

113. Evans DA, Brownlow ND, Hersh WR, Campbell EM. Automating concept identification in the electronic medical record : An experiment in extracting dosage information. Proceedings : a conference of the American Medical Informatics Association AMIA Fall Symposium. 1996 ;388–92.

114. Levin MA, Krol M, Doshi AM, Reich DL. Extraction and mapping of drug names from free text to a standardized nomenclature. AMIA Annual Symposium proceedings AMIA Symposium. 2007 Oct ;438–42.

115. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx : A medication information extraction system for clinical narratives. Journal of the American Medical Informatics Association : JAMIA [Internet]. 2010 [cited 2018 May 9] ;17(1):19–24. Available from : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2995636/

116. Hyun S, Johnson SB, Bakken S. Exploring the ability of natural language processing to extract data from nursing narratives. Computers, informatics, nursing : CIN. 2009 Jul-Aug ;27(4):215-223 ; quiz 224-225.

117. Sirohi E, Peissig P. Study of effect of drug lexicons on medication extraction from electronic medical records. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing. 2005 ;308–18.

118. Jagannathan V, Mullett CJ, Arbogast JG, Halbritter KA, Yellapragada D, Regulapati S, et al. Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. International Journal of Medical Informatics. 2009 Apr ;78(4):284–91.

119. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes : 2009 I2b2 medication extraction challenge. Journal of the American Medical Informatics Association [Internet]. 2010 Sep [cited 2020 Sep 12] ;17(5):524–7. Available from : https://academic.oup.com/jamia/article-lookup/doi/10.1136/jamia.2010.003939

120. Wei W-Q, Tao C, Jiang G, Chute CG. A High Throughput Semantic Concept Frequency Based Approach for Patient Identification : A Case Study Using Type 2 Diabetes Mellitus Clinical Notes. AMIA Annual Symposium Proceedings [Internet]. 2010 [cited 2020 Oct 16] ;2010:857–61. Available from : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041302/

121. Doan S, Collier N, Xu H, Duy PH, Phuong TM. Recognition of medication information from discharge summaries using ensembles of classifiers. BMC Medical Informatics and Decision Making [Internet]. 2012 Dec [cited 2020 Sep 12] ;12(1):36. Available from : https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-12-36

122. Li Q, Spooner SA, Kaiser M, Lingren N, Robbins J, Lingren T, et al. An end-to-end hybrid algorithm for automated medication discrepancy detection. BMC Medical Informatics and Decision Making [Internet]. 2015 Dec [cited 2020 Sep 12] ;15(1):37. Available from : http://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-015-0160-8

123. Gligic L, Kormilitzin A, Goldberg P, Nevado-Holgado A. Named entity recognition in electronic health records using transfer learning bootstrapped Neural Networks. Neural Networks : The Official Journal of the International Neural Network Society. 2020 Jan ;121:132–9.

124. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. In : Proceedings of NAACL-HLT 2016. San Diego, California, June 12-17, 2016 : ACL ; 2016. p. 260–70.

125. Wei Q, Ji Z, Li Z, Du J, Wang J, Xu J, et al. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. Journal of the American Medical Informatics Association : JAMIA. 2020 Jan ;27(1):13–21.

126. Zeng D, Sun C, Lin L, Liu B. LSTM-CRF for Drug-Named Entity Recognition. Entropy [Internet]. 2017 Jun [cited 2020 Sep 12] ;19(6):283. Available from : https://www.mdpi.com/1099-4300/19/6/283

127. Jauregi Unanue I, Zare Borzeshi E, Piccardi M. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. Journal of Biomedical Informatics. 2017 Dec ;76:102–9.

128. Li F, Liu W, Yu H. Extraction of Information Related to Adverse Drug Events from Electronic Health Record Notes : Design of an End-to-End Model Based on Deep Learning. JMIR medical informatics. 2018 Nov ;6(4):e12159.

129. Wunnava S, Qin X, Kakar T, Sen C, Rundensteiner EA, Kong X. Adverse Drug Event Detection from Electronic Health Records Using Hierarchical Recurrent Neural Networks with Dual-Level Embedding. Drug Safety. 2019 Jan ;42(1):113–22.

130. Dandala B, Joopudi V, Devarakonda M. Adverse Drug Events Detection in Clinical Notes by Jointly Modeling Entities and Relations Using Neural Networks. Drug Safety. 2019 Jan ;42(1):135–46.

131. Tao C, Filannino M, Uzuner Ö. Prescription extraction using CRFs and word embeddings. Journal of Biomedical Informatics. 2017 Aug ;72:60–6.

132. Chapman AB, Peterson KS, Alba PR, DuVall SL, Patterson OV. Detecting Adverse Drug Events with Rapidly Trained Classification Models. Drug Safety. 2019 Jan ;42(1):147–56.

133. Chen L, Gu Y, Ji X, Sun Z, Li H, Gao Y, et al. Extracting medications and associated adverse drug events using a natural language processing system combining knowledge base and deep learning. Journal of the American Medical Informatics Association [Internet]. 2020 Jan [cited 2020 Sep 12] ;27(1):56–64. Available from : https://academic.oup.com/jamia/article/27/1/56/5582687

134. Yang X, Bian J, Fang R, Bjarnadottir RI, Hogan WR, Wu Y. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. Journal of the American Medical Informatics Association [Internet]. 2020 Jan [cited 2020 Sep 12] ;27(1):65–72. Available from : https://academic.oup.com/jamia/article/27/1/65/5555856

135. Christopoulou F, Tran TT, Sahu SK, Miwa M, Ananiadou S. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. Journal of the American Medical Informatics Association [Internet]. 2020 Jan [cited 2020 Sep 12] ;27(1):39–46. Available from : https://academic.oup.com/jamia/article/27/1/39/5544735

136. Dai H-J, Su C-H, Wu C-S. Adverse drug event and medication extraction in electronic health records via a cascading architecture with different sequence labeling models and word embeddings. Journal of the American Medical Informatics Association [Internet]. 2019 Jul [cited 2020 Sep 12] ;ocz120. Available from : https://academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocz120/5537181

137. Kim Y, Meystre SM. Ensemble methodbased extraction of medication and related information from clinical texts. Journal of the American Medical Informatics Association [Internet]. 2020 Jan [cited 2020 Sep 12] ;27(1):31–8. Available from : https://academic.oup.com/jamia/article/27/1/31/5529714

138. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical Natural Language Processing in languages other than English : Opportunities and challenges. Journal of Biomedical Semantics. 2018 Mar ;9(1):12.

139. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. Journal of the American Medical Informatics Association : JAMIA. 2010 Sep-Oct ;17(5):514–8.

140. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 N2c2 shared task on adverse drug events and medication extraction in electronic health records. Journal of the American Medical Informatics Association : JAMIA. 2020 Jan ;27(1):3–12.

141. Jagannatha A, Liu F, Liu W, Yu H. Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0). Drug Safety. 2019 Jan ;42(1):99–111.

142. Johnson AEW, Pollard TJ, Shen L, Lehman L-WH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific Data. 2016 May ;3:160035.

143. Deléger L, Grouin C, Zweigenbaum P. Extracting medication information from French clinical texts. Studies in Health Technology and Informatics. 2010 ;160(Pt 2):949–53.

144. Montenegro JLZ, da Costa CA, da Rosa Righi R. Survey of conversational agents in health. Expert Systems with Applications [Internet]. 2019 Sep [cited 2020 Oct 8] ;129:56–67. Available from : http://www.sciencedirect.com/science/article/pii/S0957417419302283

145. Bordes A, Boureau Y-L, Weston J. Learning End-to-End Goal-Oriented Dialog. arXiv:160507683 [cs] [Internet]. 2016 May [cited 2017 Sep 1] ; Available from : http://arxiv.org/abs/1605.07683

146. Dong L, Mallinson J, Reddy S, Lapata M. Learning to paraphrase for question answering. In : Proceedings of the 2017 conference on empirical methods in natural language processing [Internet]. Copenhagen, Denmark : Association for Computational Linguistics ; 2017. p. 875–86. Available from : http://aclweb.org/anthology/D17-1091

147. Androutsopoulos I, Malakasiotis P. A survey of paraphrasing and textual entailment methods. Journal of Artificial Intelligence Research. 2010 ;38:135–87.

148. Madnani N, Dorr BJ. Generating phrasal and sentential paraphrases : A survey of data-driven methods. Computational Linguistics. 2010 ;36(3):341–87.

149. Iyyer M, Wieting J, Gimpel K, Zettlemoyer L. Adversarial example generation with syntactically controlled paraphrase networks. Proceedings of NAACL-HLT 2018, New Orleans, Louisiana, June 1-6, 2018. 2018 ;1875–85.

150. Ganitkevitch J, Van Durme B, Callison-Burch C. PPDB : The paraphrase database. In : Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics : Human language technologies. 2013. p. 758–64.

151. Duboue PA, Chu-Carroll J. Answering the question you wish they had asked : The impact of paraphrasing for question answering. In : Proceedings of the human language technology conference of the NAACL, companion volume : Short papers. Association for Computational Linguistics ; 2006. p. 33–6.

152. Zhang W-N, Ming Z-Y, Zhang Y, Liu T, Chua T-S. Exploring key concept paraphrasing based on pivot language translation for question retrieval. In : Proceedings of the twenty-ninth AAAI conference on artificial intelligence [Internet]. Austin, Texas : AAAI Press ; 2015. p. 410–6. (AAAI'15). Available from : http://dl.acm.org/citation.cfm?id=2887007.2887065

153. Google. Google Cloud Translation API Documentation | Translation API. Google Cloud [Internet]. Available from : https://cloud.google.com/translate/docs/

154. Wikipedia [Internet]. [cited 2018 Nov 15]. Available from : https://www.wikipedia.org/

155. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, et al. A comparison of word embeddings for the biomedical natural language processing. Journal of Biomedical Informatics [Internet]. 2018 Nov [cited 2018 Oct 5] ;87:12–20. Available from : http://www.sciencedirect.com/science/article/pii/S1532046418301825

156. Garcelon N, Neuraz A, Salomon R, Faour H, Benoit V, Delapalme A, et al. A clinician friendly data warehouse oriented toward narrative reports : Dr. Warehouse. Journal of Biomedical Informatics [Internet]. 2018 Apr [cited 2018 Jun 22] ;80:52–63. Available from : http://www.sciencedirect.com/science/article/pii/S1532046418300388

157. Weston J, Bordes A, Chopra S, Rush AM, van Merriënboer B, Joulin A, et al. Towards AI-Complete Question Answering : A Set of Prerequisite Toy Tasks. arXiv:150205698 [cs, stat] [Internet]. 2015 Feb [cited 2017 Sep 28] ; Available from : http://arxiv.org/abs/1502.05698

158. Young S, Gašić M, Thomson B, Williams JD. POMDP-Based Statistical Spoken Dialog Systems : A Review. Proceedings of the IEEE. 2013 May ;101(5):1160–79.

159. Kingma DP, Ba J. Adam : A Method for Stochastic Optimization. arXiv:14126980 [cs] [Internet]. 2017 Jan [cited 2020 Oct 12] ; Available from : http://arxiv.org/abs/1412.6980

160. Chollet F, others. Keras. 2015 ; Available from : https://github.com/fchollet/keras

161. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow : Large-Scale Machine Learning on Heterogeneous Systems. 2015 ; Available from : https://www.tensorflow.org/

162. Gupta S, Shah R, Mohit M, Kumar A, Lewis M. Semantic Parsing for Task Oriented Dialog using Hierarchical Representations. arXiv:181007942 [cs] [Internet]. 2018 Oct [cited 2020 Oct 13] ; Available from : http://arxiv.org/abs/1810.07942

163. Aghajanyan A, Maillard J, Shrivastava A, Diedrick K, Haeger M, Li H, et al. Conversational Semantic Parsing. arXiv:200913655 [cs] [Internet]. 2020 Sep [cited 2020 Oct 8] ; Available from : http://arxiv.org/abs/2009.13655

164. Dyer C, Kuncoro A, Ballesteros M, Smith NA. Recurrent Neural Network Grammars. arXiv:160207776 [cs] [Internet]. 2016 Feb [cited 2018 Oct 30] ; Available from : http://arxiv.org/abs/1602.07776

165. Sheikhshabbafghi G, Birol I, Sarkar A. In-domain Context-aware Token Embeddings Improve Biomedical Named Entity Recognition. In : Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis [Internet]. Brussels, Belgium : Association for Computational Linguistics ; 2018 [cited 2020 Oct 12]. p. 160–4. Available from : https://www.aclweb.org/anthology/W18-5618

166. Sokolov A, Filimonov D. Neural Machine Translation For Paraphrase Generation. arXiv:200614223 [cs] [Internet]. 2020 Jun [cited 2020 Oct 12] ; Available from : http://arxiv.org/abs/2006.14223

167. Kazemnejad A, Salehi M, Soleymani Baghshah M. Paraphrase Generation by Learning How to Edit from Samples. In : Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics [Internet]. Online : Association for Computational Linguistics ; 2020 [cited 2020 Oct 12]. p. 6010–21. Available from : https://www.aclweb.org/anthology/2020.acl-main.535

168. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners. arXiv:200514165 [cs] [Internet]. 2020 Jul [cited 2020 Oct 14] ; Available from : http://arxiv.org/abs/2005.14165

169. Wang P, Shi T, Reddy CK. Text-to-SQL Generation for Question Answering on Electronic Medical Records. In : Proceedings of The Web Conference 2020 [Internet]. New York, NY, USA : Association for Computing Machinery ; 2020 [cited 2020 Oct 11]. p. 350–61. (WWW '20). Available from : https://doi.org/10.1145/3366423.3380120

170. Daniel C, Serre P, Orlova N, Bréant S, Paris N, Griffon N. Initializing a hospital-wide data quality program. The AP-HP experience. Computer Methods and Programs in Biomedicine [Internet]. 2019 Nov [cited 2020 Oct 3] ;181:104804. Available from : http://www.sciencedirect.com/science/article/pii/S0169260718306242

171. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (I2b2). J Am Med Inform Assoc [Internet]. 2010 Mar ;17:124–30. Available from : http://www.ncbi.nlm.nih.gov/pubmed/20190053%20http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3000779/pdf/jamia000893.pdf

172. Sang EFTK, De Meulder F. Introduction to the CoNLL-2003 Shared Task : Language-Independent Named Entity Recognition. arXiv:cs/0306050 [Internet]. 2003 Jun [cited 2020 Oct 3] ; Available from : http://arxiv.org/abs/cs/0306050

173. Akbik A, Bergmann T, Blythe D, Rasul K, Schweter S, Vollgraf R. FLAIR : An Easy-to-Use Framework for State-of-the-Art NLP. In : Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations) [Internet]. Minneapolis, Minnesota : Association for Computational Linguistics ; 2019 [cited 2020 Jun 30]. p. 54–9. Available from : https://www.aclweb.org/anthology/N19-4010

174. Polyak BT, Juditsky AB. Acceleration of Stochastic Approximation by Averaging. SIAM Journal on Control and Optimization [Internet]. 1992 Jul [cited 2020 Oct 4] ;30(4):838–55. Available from : https://epubs.siam.org/doi/abs/10.1137/0330046

175. Agarwal K, Eftimov T, Addanki R, Choudhury S, Tamang S, Rallo R. Snomed2Vec : Random Walk and Poincar\'e Embeddings of a Clinical Knowledge Base for Healthcare Analytics. arXiv:190708650 [cs, stat] [Internet]. 2019 Jul [cited 2020 Oct 7] ; Available from : http://arxiv.org/abs/1907.08650

176. Pyysalo S, Ohta T, Rak R, Sullivan D, Mao C, Wang C, et al. Overview of the ID, EPI and REL tasks of BioNLP Shared Task 2011. BMC Bioinformatics [Internet]. 2012 Jun [cited 2020 Oct 12] ;13(11):S2. Available from : https://doi.org/10.1186/1471-2105-13-S11-S2

177. Comeau DC, Islamaj Doğan R, Ciccarese P, Cohen KB, Krallinger M, Leitner F, et al. BioC : A minimalist approach to interoperability for biomedical text processing. Database : The Journal of Biological Databases and Curation. 2013 ;2013:bat064.

178. Bender D, Sartipi K. HL7 FHIR : An Agile and RESTful approach to healthcare information exchange. In : Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems. 2013. p. 326–31.

179. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI) : Opportunities for Observational Researchers. Studies in Health Technology and Informatics. 2015 ;216:574–8.

180. Ferrucci D, Lally A. UIMA : An architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering [Internet]. 2004 Sep [cited 2020 Oct 12] ;10(3-4):327–48. Available from : https://doi.org/10.1017/S1351324904003523

181. Carrell DS, Schoen RE, Leffler DA, Morris M, Rose S, Baer A, et al. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. Journal of the American Medical Informatics Association [Internet]. 2017 Sep [cited 2020 Oct 12] ;24(5):986–91. Available from : https://academic.oup.com/jamia/article/24/5/986/3737804

182. Chapman WW, Dowling JN, Ivanov O, Gesteland PH, Olszewski R, Espino JU, et al. Evaluating natural language processing applications applied to outbreak and disease surveillance. In : Proceedings of 36th symposium on the interface : Computing science and statistics. Citeseer ; 2004.

183. Elkin PL, Froehling DA, Wahner-Roedler DL, Brown SH, Bailey KR. Comparison of natural language processing biosurveillance methods for identifying influenza from encounter notes. Annals of Internal Medicine. 2012 ;156(1_Part_1):11–8.

184. Zhang L, Sun Y, Zeng H-L, Peng Y, Jiang X, Shang W-J, et al. Calcium channel blocker amlodipine besylate is associated with reduced case fatality rate of COVID-19 patients with hypertension. medRxiv [Internet]. 2020 Apr [cited 2020 Apr 24] ;2020.04.08.20047134. Available from : https://www.medrxiv.org/content/10.1101/2020.04.08.20047134v1

185. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Archive [Internet]. 1993 [cited 2017 May 20] ;32:281–91. Available from : https://methods.schattauer.de/en/contents/archivepremium/issue/1198/manuscript/14376.html

186. Okazaki N, Tsujii J. Simple and Efficient Algorithm for Approximate Dictionary Matching. In : Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010) [Internet]. Beijing, China : Coling 2010 Organizing Committee ; 2010 [cited 2020 May 8]. p. 851–9. Available from : https://www.aclweb.org/anthology/C10-1096

187. Garcelon N, Neuraz A, Benoit V, Salomon R, Burgun A. Improving a full-text search engine : The importance of negation detection and family history context to identify cases in a biomedical data warehouse. Journal of the American Medical Informatics Association : JAMIA. 2017 May ;24(3):607–13.

188. Cossin S, Lebrun L, Lobre G, Loustau R, Jouhet V, Griffier R, et al. Romedi : An Open Data Source About French Drugs on the Semantic Web. Studies in Health Technology and Informatics. 2019 Aug ;264:79–82.

189. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's transformers : State-of-the-art natural language processing. ArXiv. 2019 ;abs/1910.03771.

190. Olivier M. Modulation of host cell intracellular Ca2+. Parasitology Today [Internet]. 1996 Apr [cited 2020 Aug 19] ;12(4):145–50. Available from : http://www.sciencedirect.com/science/article/pii/0169475896100065

191. Muthuswamy Balasubramanyam. COVID-19 : Is it time to revisit the research on calcium channel drug targets ? [Internet]. European Medical Journal. 2020 [cited 2020 Aug 19]. Available from : https://www.emjreviews.com/diabetes/article/covid-19-is-it-time-to-revisit-the-research-on-calcium-channel-drug-targets/

192. R Core Team. R : A language and environment for statistical computing [Internet]. Vienna, Austria : R Foundation for Statistical Computing ; 2020. Available from : https://www.R-project.org/

193. Therneau TM. A package for survival analysis in s [Internet]. 2015. Available from : https://CRAN.R-project.org/package=survival

194. European Medicines Agency. Strategic reflection. EMA Regulatory Science to 2025. 2018 ;79.

195. Hoertel N, Rico MS, Vernet R, Beeker N, Jannot A-S, Neuraz A, et al. Association between SSRI Antidepressant Use and Reduced Risk of Intubation or Death in Hospitalized Patients with Coronavirus Disease 2019 : A Multicenter Retrospective Observational Study. medRxiv [Internet]. 2020 Aug [cited 2020 Sep 1] ;2020.07.09.20143339. Available from : https://www.medrxiv.org/content/10.1101/2020.07.09.20143339v2

196. Hoertel N, Rico MS, Vernet R, Jannot A-S, Neuraz A, Blanco C, et al. Observational Study of Haloperidol in Hospitalized Patients with Covid-19. medRxiv [Internet]. 2020 Jul [cited 2020 Sep 1] ;2020.07.15.20150490. Available from : https://www.medrxiv.org/content/10.1101/2020.07.15.20150490v1

197. Sbidian E, Josse J, Lemaitre G, Mayer I, Bernaux M, Gramfort A, et al. Hydroxychloroquine with or without azithromycin and in-hospital mortality or discharge in patients hospitalized for COVID-19 infection : A cohort study of 4,642 in-patients in France. medRxiv [Internet]. 2020 Jun [cited 2020 Sep 1] ;2020.06.16.20132597. Available from : https://www.medrxiv.org/content/10.1101/2020.06.16.20132597v1

198. Pizer SD. Falsification Testing of Instrumental Variables Methods for Comparative Effectiveness Research. Health Services Research [Internet]. 2016 [cited 2020 Jun 30] ;51(2):790–811. Available from : https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-6773.12355

199. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT : A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics [Internet]. 2020 Feb [cited 2020 Oct 16] ;36(4):1234–40. Available from : https://academic.oup.com/bioinformatics/article/36/4/1234/5566506

200. Du J, Grave E, Gunel B, Chaudhary V, Celebi O, Auli M, et al. Self-training Improves Pre-training for Natural Language Understanding. arXiv:201002194 [cs] [Internet]. 2020 Oct [cited 2020 Oct 19] ; Available from : http://arxiv.org/abs/2010.02194

201. Huang S-C, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning : A systematic review and implementation guidelines. npj Digital Medicine [Internet]. 2020 Oct [cited 2020 Oct 19] ;3(1):1–9. Available from : https://www.nature.com/articles/s41746-020-00341-z

202. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient : An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Scientific Reports [Internet]. 2016 May [cited 2017 Apr 25] ;6:26094. Available from : http://www.nature.com.gate2.inist.fr/srep/2016/160517/srep26094/full/srep26094.html

203. Hoertel N, Rico MS, Vernet R, Jannot A-S, Neuraz A, Blanco C, et al. Observational Study of Chlorpromazine in Hospitalized Patients with Covid-19. medRxiv [Internet]. 2020 Jul [cited 2020 Oct 19] ;2020.07.15.20154310. Available from : https://www.medrxiv.org/content/10.1101/2020.07.15.20154310v1

204. Brat GA, Weber GM, Gehlenborg N, Avillach P, Palmer NP, Chiovato L, et al. International electronic health record-derived COVID-19 clinical course profiles : The 4CE consortium. npj Digital Medicine [Internet]. 2020 Aug [cited 2020 Oct 19] ;3(1):1–9. Available from : https://www.nature.com/articles/s41746-020-00308-0

Université de Paris

Titre : Compréhension du langage naturel pour le dossier patient informatisé

Mots Clefs : traitement automatique de la langue ; réseaux de neurones profond ; dossier patient informatisé ; extraction d'information, extraction d'entités nommées ; compréhension de la langue

Résumé : Dans le domaine médical, la langue naturelle tient une place particulièrement importante pour la communication et le stockage d'informations. En effet, outre les données dites "structurées" (*e.g.*, les résultats d'examens biologiques), la langue naturelle est omniprésente : formulaires de demande d'examens, notes de suivi clinique, comptes-rendus d'hospitalisation, comptes-rendus d'examens d'imagerie, en sont des exemples. Ce langage naturel médical est complexe et difficile à maîtriser : il faut plusieurs années aux futurs médecins pour apprendre à le déchiffrer correctement. En effet, le jargon y est omniprésent, ainsi que des références à des connaissances implicites, des abréviations inconstantes ou encore des fautes d'orthographe ou de frappe. Malgré la difficulté, entraîner des machines à comprendre le texte médical, soit pour faciliter l'accès à l'information, soit pour extraire de l'information, est une tâche essentielle pour améliorer à la fois l'accès à l'information et les connaissances médicales. La première partie de cette thèse concerne l'accès aux informations et s'intéresse à la compréhension du langage naturel dans le cadre d'un agent conversationnel permettant d'interroger le dossier patient informatisé. Nous nous sommes intéressés à des techniques de supervision distante (*i.e.*, génération, paraphrase) pour entraîner un modèle de compréhension de la langue en l'absence de données d'entraînement basé sur des réseaux de neurones récurrents. Nous avons également étudié l'apport de plongements lexicaux contextualisés (word embeddings) spécialisés sur des tâches de compréhension du langage médical. Dans la deuxième partie, nous nous sommes intéressés à l'extraction d'informations sur les médicaments dans les textes clinique. Nous avons en premier lieu développé un corpus de textes cliniques annotés, et un modèle d'extraction hybride combinant règles expertes et apprentissage par réseaux de neurones récurrents. Par la suite, nous avons montré l'intérêt de déployer de tels systèmes à grande échelle pour assurer une réponse rapide dans le cadre de maladies émergentes telles que la COVID-19.

Title : Natural language understanding for the electronic health records : access to information and information extraction

Keys words : natural language processing ; deep learning, electronic health record ; information extraction ; natural language understanding ; named entity recognition

Abstract : In the medical field, natural language plays an important role in communication and information storage. Indeed, in addition to structured data (*e.g.*, results of biological tests), natural language is omnipresent : discharge summaries, clinical follow-up notes, hospitalization reports, radiologic tests results are examples of this. This natural medical language is complex and difficult to master : it takes several years for future doctors to learn how to decipher it correctly. Indeed, jargon is omnipresent, as well as references to implicit knowledge, inconsistent abbreviations, spelling and typing errors. Despite the difficulty, training machines to understand medical text, either to facilitate access to information or to extract information, is an essential task to improve both access to information and medical knowledge. A first part of this thesis deals with access to information and focuses on the understanding of natural language in the context of a conversational agent allowing to query the computerized patient record. We leveraged in distant supervision techniques (*i.e.*, generation, paraphrase) to train a model of language comprehension in the absence of training data, based on recurrent neural networks. We have also studied the contribution of specialized contextualized word embeddings on medical language comprehension tasks. In the second part, we focused on the extraction of drug information from clinical texts. We first developed a corpus of annotated clinical texts, and a hybrid extraction model combining expert rules and recurrent neural networks. Subsequently, we showed the interest of deploying such systems at a large scale to provide a rapid response in the context of emerging diseases such as COVID-19.