Adverse Drug Reaction Detection

The Data Science at the service of PharmacoVigilance

EDSON FLOREZ, PHD STUDENT
FREDERIC PRECIOSO, MICHEL RIVEILL (I3S)
DAVID DARMON, PASCAL STACCINI (CHU DE NICE)
Adverse Drug Reaction (ADR) detection

Problem of detect correlations in the medical research domain, to establish if an adverse clinical state have appeared after taking medicines.

**Input:** Clinical notes contain medication, disorders, diseases, etc.

**Output:** Adverse Drug Reaction (ADR) of treatments

“the patient has internal bleeding secondary to warfarin”

Disorder[DISO]>>> ADR >>> Medication[CHEM]

ADR is an adverse event caused by a drug (Relation)
1. Preprocessing
   - 1. Spelling corrector
   - 2. Abbreviations and Acronyms
   - 3. Lemmatization

2. Annotation
   - Entity Recognition
     Identifying labels CHEM, DISO, etc

3. ADR detection
   - Relation Extraction
     Identification of Drug-DISO relation
   - Transform some DISO in ADR

Clinical Text

Ontologies
## PRIMEGE Database

Collect anonymous data directly from the consultation software (EHR), with no effort of doctors to feed the database for research purpose.

Data of 13 GPs (general practitioners) collected from 2012

Data contain both structured text with codes and notes in free text (unstructured)

### Table: Volume de données

<table>
<thead>
<tr>
<th>Elément</th>
<th>Nombre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients</td>
<td>38 970</td>
</tr>
<tr>
<td>Consultations</td>
<td>241 472</td>
</tr>
<tr>
<td>Antécédents</td>
<td>146 333</td>
</tr>
<tr>
<td>Données de biométrie</td>
<td>218 371</td>
</tr>
<tr>
<td>Motifs de consultation</td>
<td>157 009</td>
</tr>
<tr>
<td>Diagnostics</td>
<td>109 181</td>
</tr>
<tr>
<td>Lignes de prescriptions médicamenteuses</td>
<td>560 536</td>
</tr>
<tr>
<td>Symptômes recueillis</td>
<td>11 638</td>
</tr>
<tr>
<td>Procédures de soin</td>
<td>8 146</td>
</tr>
<tr>
<td>Examen complémentaires</td>
<td>546 262</td>
</tr>
<tr>
<td>Prescriptions paramédicales</td>
<td>11 048</td>
</tr>
<tr>
<td>Observations/notes</td>
<td>36 702</td>
</tr>
</tbody>
</table>

Data until 2016

- medical observation and symptoms
- weight, blood pressure ....
- diagnoses
- reasons of encounter
- radiology results
- data from another doctor
- Adverse Drug Event
1. Pre-processing

Denoising of PRIMEGE notes

- Abbreviations and medical jargon
  - cardio instead: cardiologue cardiologique cardiomégalie cardiopathie
  - gastro instead: gastroenterite gastroenterologue gastroscopie gastrocnemiens
  - pulm instead pulmonaire, cardiopulm, cardiopulmonaire
  - nl=normal, g. =gauche, qq=quelques, dte=droite, trt=traitement, tr=trouble

- Correction of wrong spelling words
  - apetit instead appétit
  - esport instead sport

- Medical Acronyms
  - TA, AB, TDR, ADP, ASD, MT, RC, EFR, etc
2. Annotation

Extracting Concepts with Multiple Terminologies (ECMT) tool** by CISMef* (Rouen CHU), which annotates text using the concepts of health ontologies/terminologies in French.

- **Terme**: preferred Term (for synonyms)
- **Ter**: Terminology acronym: 55 terminologies included in the repository HeTOP [11]
- **Code**: Internal code of the terminology
- **CUI**: Concept UMLS Identifier: Unified Medical Language System (UMLS) code


*CISMef Catalog and Index of French language Medical Sites. Available in: http://www.chu-rouen.fr/cismef/

WHO-ART Ontology

WHO-ART : specialized ontology on Adverse Reaction Terminology

Example: searching **Fievre** (Fever) within WHO-ART with the annotator, in french or english

**WHO-ART** ontology (Main Class)

- **ETAT GENERAL** (Class)  => **DISOrder** (High level label)
  - **FIEVRE** (Preferred Term), Term included (Synonyms):
    - **FIEVRE D'ORIGINE MEDICAMENTEUSE**
    - **PYREXIE**
    - **REACTION FEBRILE**


ECMT tool* extracted ontology concepts on PRIMEGE observations.csv

Basic annotation: [Begin, End, Token, Label, #Annotation, Ontology, Preferred Term]

Line 17: d+ nuque et au dessus de l'oreille, mieux avec advil, surveillance 3J. oreille nl -
[[49,54,"advil","CHEM",1,"PHA","ADVIL"]]

Line 18: se plaint de fluctuation d'anxiété -
[[27,34,"anxiété","PHYS",1,"ART","ANXIETE"]]

Line 19: rhinorrhée post, apyrexie pas de perte d'appétit - surveillance -
[[0,9,"rhinorrhée","DISO",1,"ART","RHINORRHEE"],
[33,47,"perte d'appétit","DISO",2,"ICD","anorexie"]

Term >> Preferred Term (synonym)

ECMT annotations of PRIMEGE notes

ECMT tool* extracted ontology concepts on PRIMEGE: 46422 notes with 1413030 tokens

<table>
<thead>
<tr>
<th>Ontology</th>
<th>Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>ART</td>
<td>34334</td>
</tr>
<tr>
<td>PHA</td>
<td>24225</td>
</tr>
<tr>
<td>ICD</td>
<td>19451</td>
</tr>
<tr>
<td>Overall</td>
<td>78010</td>
</tr>
</tbody>
</table>

- 3 Ontologies:  
  - WHO-ART (Adverse Reaction Terminology)  
  - PHA (Racine des médicaments)  
  - ICD-10 (International Statistical Classification of Diseases)

- 7 categories (Label): ACTivities and Behaviors, CHEMical and drugs, DISOrders, PHENomena, PHYSiology, PROCedures, CONCepts and Ideas

<table>
<thead>
<tr>
<th>Label</th>
<th>Annotations</th>
<th>Ontology</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTI</td>
<td>272</td>
<td>Common</td>
</tr>
<tr>
<td>CHEM</td>
<td>24225</td>
<td>PHA</td>
</tr>
<tr>
<td>DISO</td>
<td>52403</td>
<td>Common</td>
</tr>
<tr>
<td>PHEN</td>
<td>295</td>
<td>Common</td>
</tr>
<tr>
<td>PHYS</td>
<td>796</td>
<td>Common</td>
</tr>
<tr>
<td>PROC</td>
<td>17</td>
<td>ICD</td>
</tr>
<tr>
<td>CONC</td>
<td>2</td>
<td>ICD</td>
</tr>
<tr>
<td>Overall</td>
<td>78010</td>
<td>5.5% of total tokens</td>
</tr>
</tbody>
</table>

CISMEF code: e.g. ART_IT_0171_IT5, DISO label for term “sensation bizarre”

3. Extracting ADR relation in Clinical Notes

Detection of relation between ADR and Medication in the sentence:

“The patient has internal bleeding (A) secondary to warfarin (B)”

[ADR]>>> dependency >>> [CHEM] Medication

Label for entity A is strongly related to the label prediction of B.

RNN (Recurrent Neural Network) models as LSTM (Long Short Term Memory) can model long term label dependencies [1].

3. Extracting ADR relation in Clinical Notes with LSTM

**Dataset**
Text + Annotations

1. Training

**Learning algorithm**
LSTM network

2. Test

**Model**

**Output**

Labels:
- None
- CHEM
- DISO
- ADR

**Input**

*next Sentences*

The patient has *internal bleeding* [ADR] secondary to *warfarin* [CHEM].

*For the treatment of fever* [DISO], the patient took an *aspirin* [CHEM].
PRIMEGE database

Notes (id) are linked with the other elements of the database

Entity Association model centered on the visits*

STATUS and FUTURE WORK

1. *First version of tool to extract clinical entities like disorders, medications, etc.*

2. *Obtain for each patient a time series with:*
   - Notes
   - Medication
   - Diagnostic
   - Symptoms, etc

3. *Build and train LSTM in order to detect ADR*
# Named Entity Recognition approaches

Comparison of system for entity recognition when manually annotated data (from different sources) is used for evaluation [9]

<table>
<thead>
<tr>
<th>Study</th>
<th>Reference</th>
<th>Method</th>
<th>Size</th>
<th>Recall</th>
<th>Precision</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaman</td>
<td>[17]**</td>
<td>Lexicon-based (450 comments for system development)</td>
<td>3150</td>
<td>0.70</td>
<td>0.78</td>
<td>0.74</td>
</tr>
<tr>
<td>Nikfarjam</td>
<td>[8]**</td>
<td>Lexical pattern-matching (2400 comments). Association rule mining to identify patterns</td>
<td>1200</td>
<td>0.66</td>
<td>0.70</td>
<td>0.68</td>
</tr>
<tr>
<td>Nikfarjam</td>
<td>[3]</td>
<td>Supervised learning via Conditional Random Fields (CRFs)</td>
<td>1559</td>
<td>0.78</td>
<td>0.86</td>
<td>0.82</td>
</tr>
<tr>
<td>Jagannatha</td>
<td>[1]</td>
<td>Bi-LSTM-CRF (skip-crf-approx)</td>
<td>1154</td>
<td>0.83</td>
<td>0.81</td>
<td>0.82</td>
</tr>
<tr>
<td>Huynh</td>
<td>[5]*</td>
<td>CNNA (Convolutional Neural Network)</td>
<td>2972</td>
<td>0.84</td>
<td>0.82</td>
<td>0.83</td>
</tr>
<tr>
<td>Gurulingappa</td>
<td>[10]*</td>
<td>SVM (Support Vector Machines)</td>
<td>2972</td>
<td>0.86</td>
<td>0.89</td>
<td>0.87</td>
</tr>
</tbody>
</table>

*Systems using the same dataset or **subsets of the same dataset