Automated Drug-Related Information Extraction from French Clinical Documents: ReLyfe Approach

Dr. Azzam Alwan, azzam.alwan@relyfe.com
Maayane Attias, maayane-lea.attias@polytechnique.edu
Dr. Larry Rubin, larry.rubin@becarelink.com
Dr. Adnan El Bakri, ceo@relyfe.com

ReLyfe R&D departement

October 03, 2021 - Barcelona, Spain

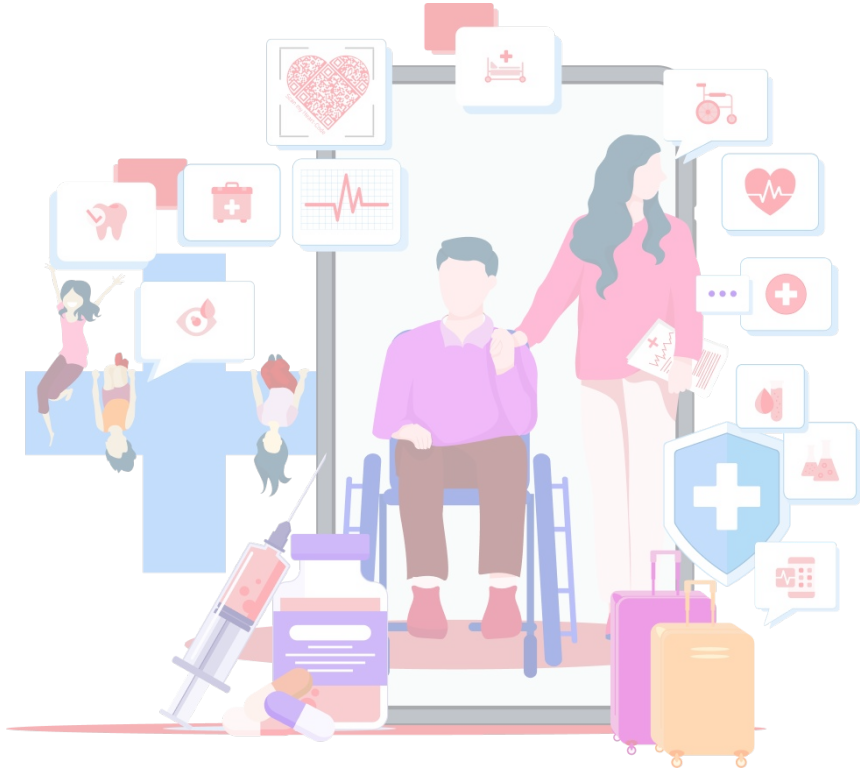
The Tenth International Conference
on Global Health Challenges
GLOBAL HEALTH 2021

presenter

Azzam Alwan received the Diploma and M.S. degrees from the Grenoble Polytechnic School and the Ph.D. degree in systems optimization and safety from the University of Technology of Troyes, in 2019. He is currently a Senior Research Engineer with Relyfe.com. In 2020, he was considered among the top eight innovators in the Middle East and Africa region due to his innovation of a new deep learning architecture to recognize the behavior of smartphone users. His research interests include chronic diseases prediction and medical document analysis.

Plan

- Problematic
- State-of-the-art
- Materials and Methods
- Evaluation and Results
- Conclusion



Problematic

- **Problems :**
  - Lack of information doctors have about their patients.
  - Transferred documents from one doctor to another.
  - Wrong actions that are incoherent with the patient’s status.
  - Incompatibility between medication and their medical condition.
  - Healthcare systems do not have patient’s information gathered all in one place.
  - 80% of relevant clinical information exist only in an unstructured form.

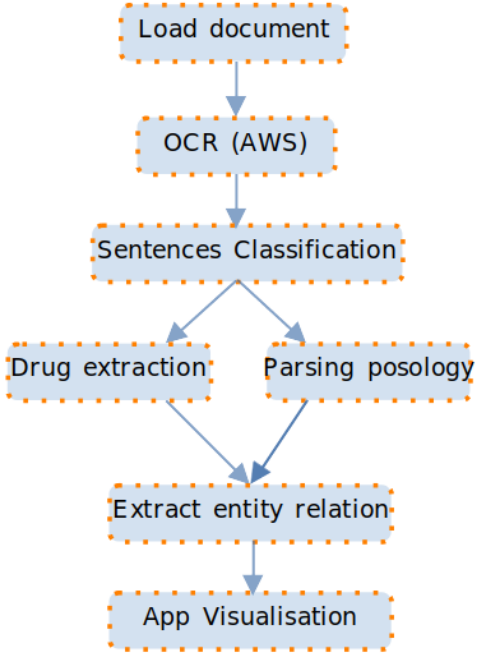
- **Solutions:**
  - Automatic Structuring of French medical prescriptions.
  - Clinical Natural Language Processing (NLP) and Named Entity Recongition (NER) model for the relevant medical information.

State-of-the-art

- **Related work:**
  - The majority of NLP research work on medical data has been carried out on texts/documents in English.
  - Since 2010, there have been less than five pertinent papers talking about the extraction of drug-related information from French clinical prescriptions.
  - Lexicon-based approaches use predefined lexicons or regular expressions to match parts of the text to recognize predefined entities (MedEx and MedXN).
  - NER model relying on a bidirectional long-short-term memory with conditional random fields (BiLSTM-CRF).
  - A conditional random field for the NER model along with a support vector machine extracting related entities combined with a rule-based context engine.

Materials and Methods

- **Method**
  - 5 main steps :
    - Optical Character Recognition (OCR).
    - Sentences classification model to differentiate between Drug, posology and private expressions.
    - Applying the two NER models (NER for the drug name, and the NER for its related information).
    - Connecting the drug to its related posolgy.
    - Finaly the display on our web application.



Materials and Methods

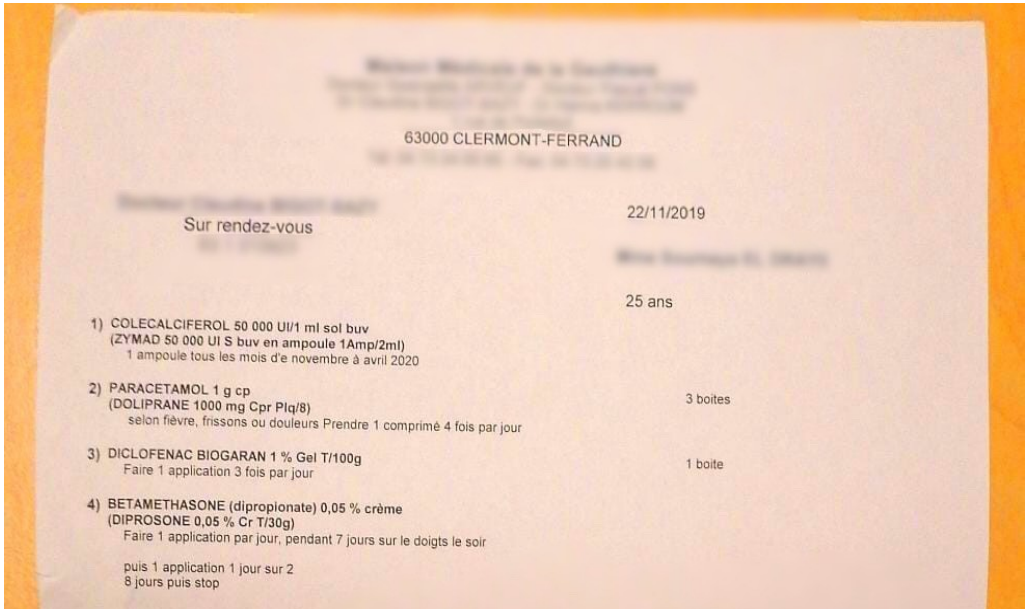
- **Data**
  - **Data for the NER models :**
    - 90% of the French prescription could be categorized into three types of formats.
    - It consists of : 100 prescriptions from 10 different cities and different clinics from 47 persons.
    - Each of them has at least five drug names followed by 1-3 sentences describing how it should be taken (posology).
  - **Data for the sentence classification model :**
    - 15000 drug names from the government public drug database.
    - We have generated 15000 synthetic posologies.
    - 15000 sentences carrying medical information, patient information, names, and all kinds of information that may be present in a prescription other than the drug and the posology.

Materials and Methods

- **Annotation Tools**
  - Prodigy: web application developed by Spacy library team.
  - It is a treaming display of all the document sentences one by one. The data scientist's role is to choose one of the predefined categories for each sentence as a label.
- **OCR**
  - We used the Aws service « Textract » to extract the text from the scanned prescriptions.
  - Documents were anonymized before being treated.
- **Text Pre-processing**
  - Unify the words format (remove the accents, lowercase, remove spaces, stop words,...)

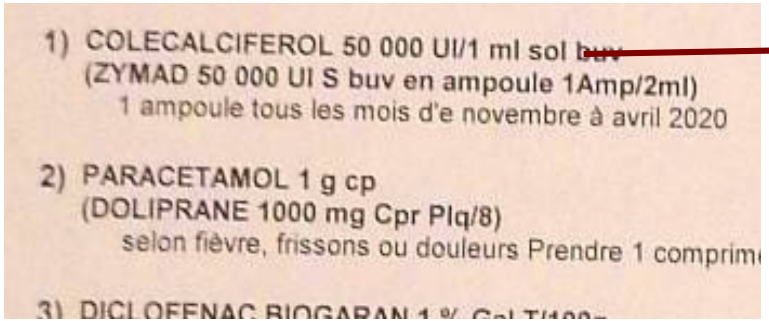
Materials and Methods

- **Sentences classification**
  - LSTM-Bi architecture is used to classify sentences into three categories :
    drug, posology, and useless sentences.
  - 95.23% Accuracy.



Materials and Methods

- **Drug detection**
  - Drug-matcher relied on a rule-based approach using the French Government drug databases.
  - Attach the detected drug to a unique international ID via the Vidal database.
  - If the first drug name does not exist in Vidal, we look at the second which is in parentheses.



Vidal output

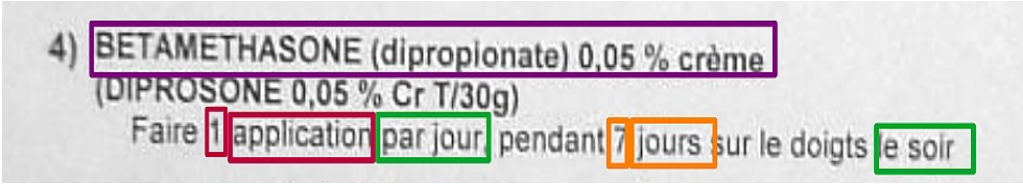

```
▸ vidal:categories : Array[1] ["VMP"]
▾ title :
    0 : "COLECALCIFEROL 50 000 UI/2 ml sol buv"
▸ link : Array[17] [{"rel":["alternate"],"type":["ap
▸ category : Array[1] [{"term":["VMP"]}]
▸ author : Array[1] [{"name":["VIDAL"]}]
▾ id :
    0 : "vidal://vmp/18301"
▸ updated : Array[1] ["2021-07-20T00:00:00Z"]
▸ summary : Array[1] [{"_":"colécalciférol * 50 000
▸ content : Array[1] [""]
```

Materials and Methods

- **Posology detection**
  - We employed the rule-based model from the NLP Spacy library to create the entity matcher.
  - We designed four matchers for the four different features we are looking for (dosage, frequency, duration, comment).

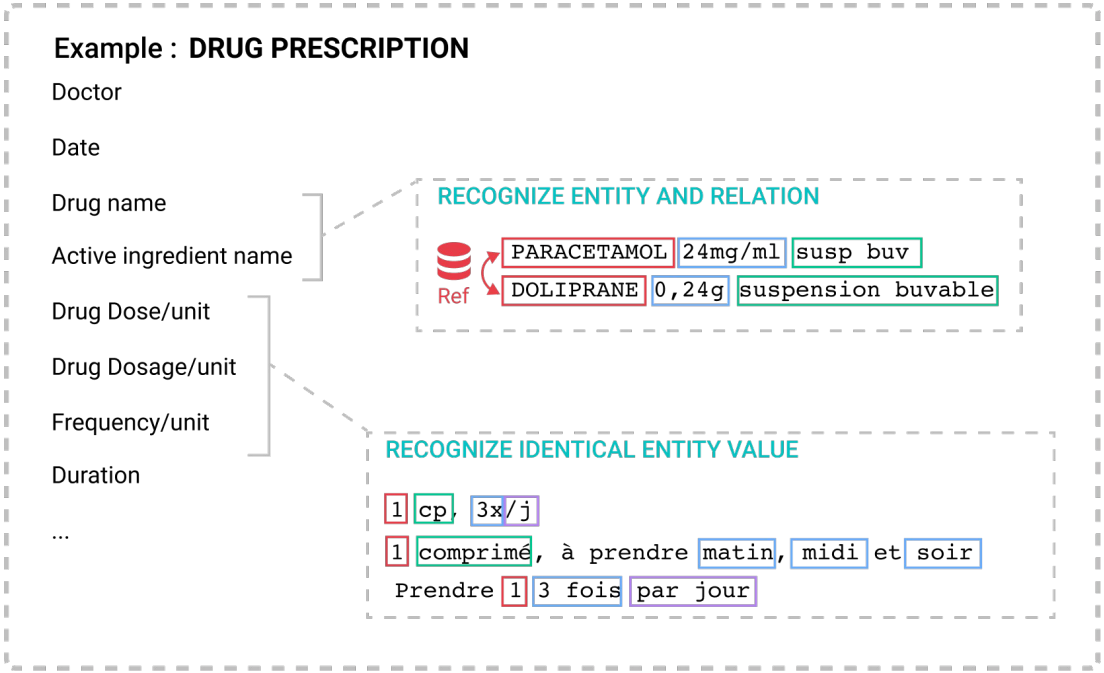
```
{"label":"DOSE","pattern":[{"LIKE_NUM": True}, {"LOWER" : {"REGEX":"(graduation[s]?)"}}]}
{"label":"DOSE","pattern":[{"LOWER": {"REGEX":"([\d]amp)"}, '_': {'position_token':True}}]}
{"label":"DOSE","pattern":[{"LIKE_NUM": True}, {"LOWER": {"REGEX":"(ampoule[s]?)"}}]}
{"label":"DOSE","pattern":[{"LOWER": "une"}, {"LOWER": {"REGEX":"(ampoule[s]?)"}}]}
```

rules created for the "dose matcher"



Materials and Methods

- **Drug-posology relation extraction**



Evaluation and Reuslts

- **Evaluation**
  - F1-Score :
    - An entity is considered a true-positive when it was annotated with the correct label.
    - False-positive when a token is falsely annotated with respect to each feature.
    - False-negative when it was not annotated at all, or it was annotated with the incorrect label.
- **Results**

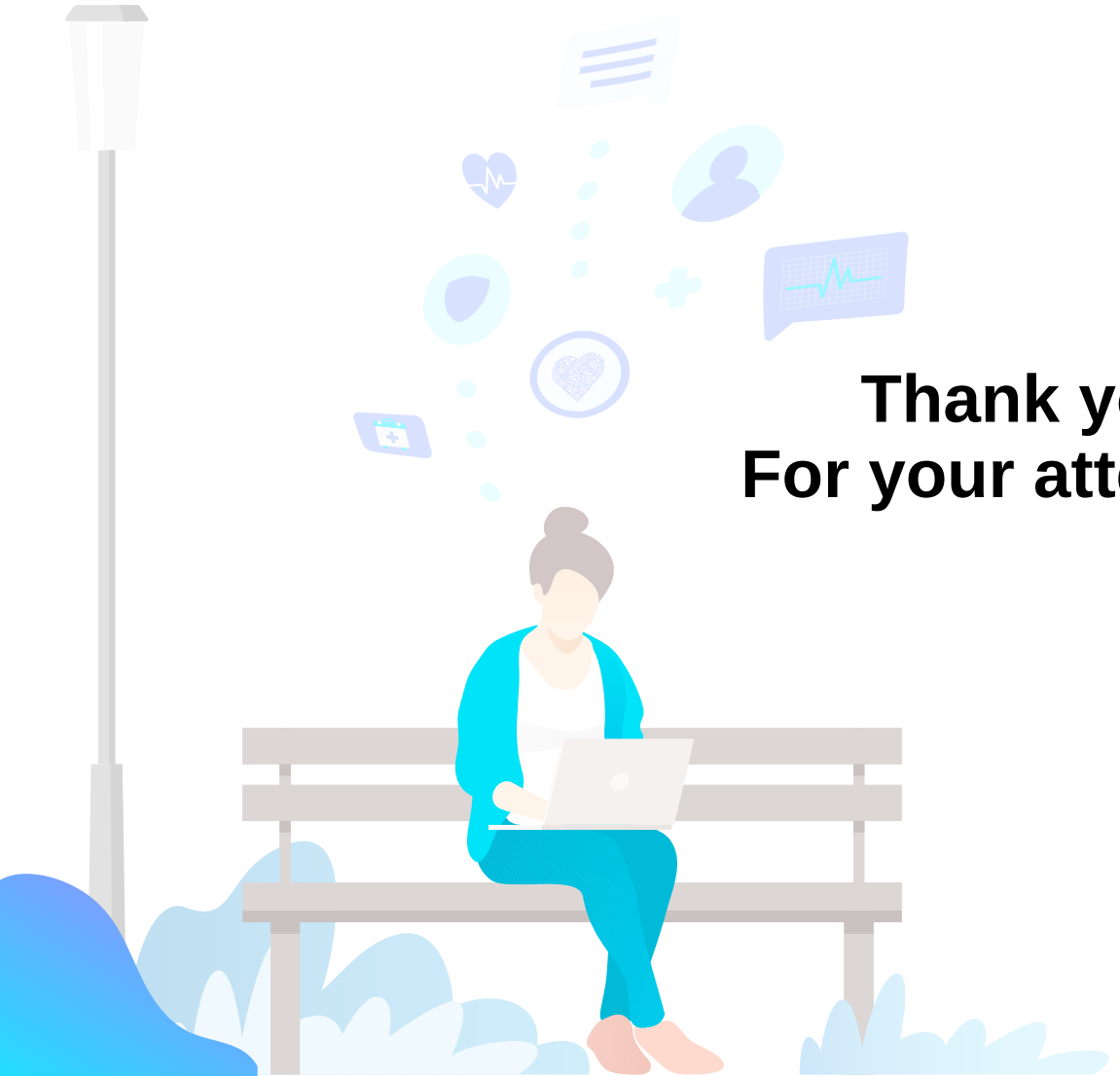
Label	F-measure	Precision	Recall
Drug name	94.33	100	89.33
Dose	93.91	100	88.52
Duration	94.91	98.24	91.80
Frequency	96.60	100	93.44
Comment	91.10	100	83.60

Evaluation and Reuslts

- **Results**
  - Add scanned document as a new treatment to your account.
  - Display the extracted entities in a web application in a very nice way and therefore will be useful for the user and the doctor.

Conclusion

- Real-time application to extract drug-related information.
- This work is one of the few that contributes to the French medical documents.
- Our approach is a series of methods concatenated together to achieve a high-performance system capable of coping with the constraints of real applications.
- Our algorithm based on human intuition and the sentence's geometric position.
- Theoretical and practical tests have proved the outperformance of our approach.
- Future Work :
    - Create a model based on a combination of Word embedding and a rule-based approach.



Thank you
For your attention