FRASIMED: a Clinical French Annotated Resource Produced through Crosslingual BERT-Based Annotation Projection

Jamil Zaghir, Mina Bjelogrlic, Jean-Philippe Goldman, Soukaïna Aananou, Christophe Gaudet-Blavignac, Christian Lovis

Crosslingual Annotation Projection

Challenge

Manual annotation is expensive

Question

Leverage existing annotations?

<u>Build your new dataset</u> from datasets in other languages

We release **a pipeline** to do **crosslingual annotation projection** in **more than 100 languages** while **preserving entity links**.

Try a <u>new French clinical dataset</u> with knowledge links

We release **FRASIMED**, a large **French** annotated dataset which contains synthetic **clinical cases**. It comprises:
- **24'037 annotations**
- **20'589 SNOMED-CT** links
- **15'457 ICD-O** links

Release to the Public

Automatic pipeline

Language independent Python code working for **most bilingual corpora**

GitHub repository:

FRASIMED

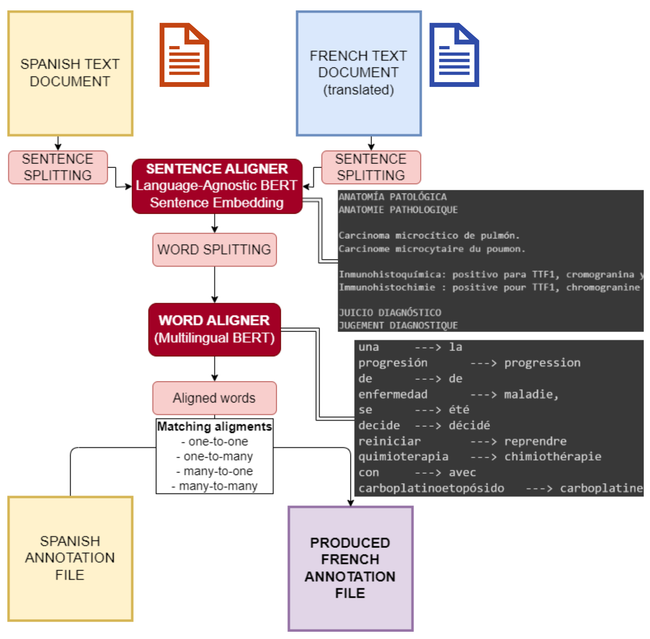
FRASIMED

Largest French clinical dataset with **oncological annotations** and **links to SNOMED-CT**

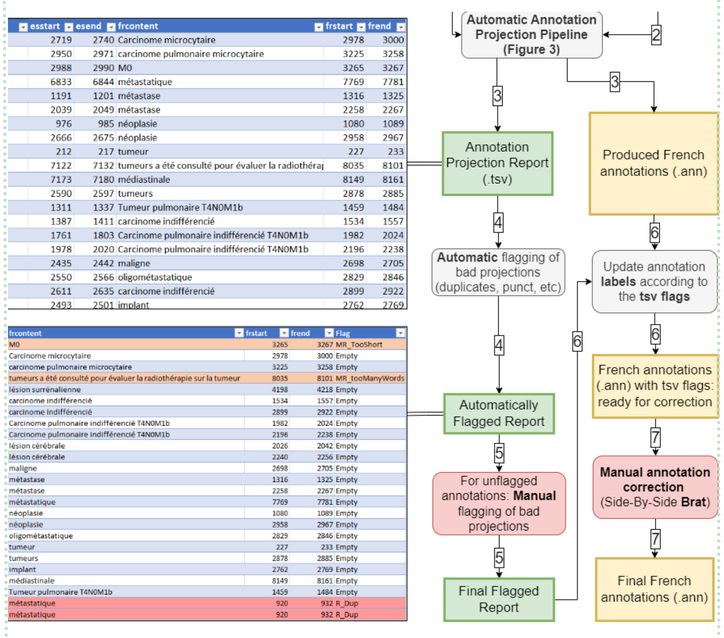
Zenodo repository:

Methodology

1- Automatic Projection Pipeline



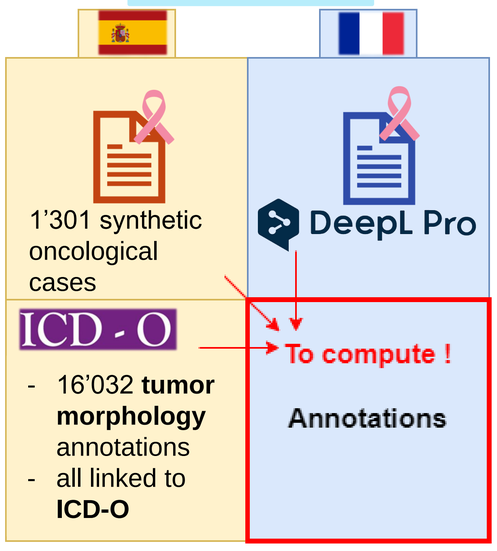
2- Semi-Automatic Error Detection



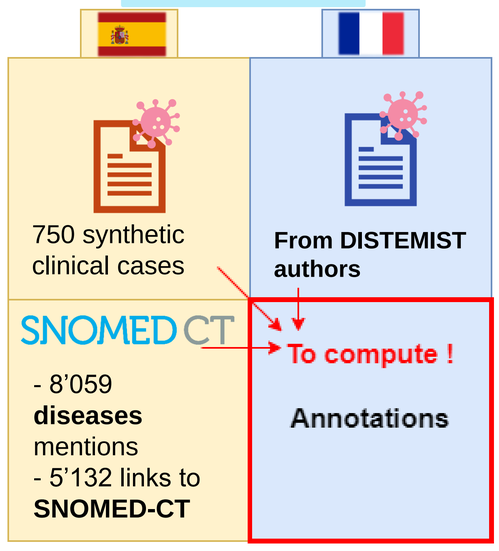
3- Side-by-side Correction

Datasets

CANTEMIST [1]



DISTEMIST [2]



Evaluation

Qualitative

640 projection pairs manually evaluated by a **clinician expert**

Error rate: **1.4%**

Quantitative

Dataset	# Entities	# Codes
CANTEMIST-ES (Miranda-Escalada et al. 2020)	16'032	16'032
CANTEMIST-FR (ours)	15'978	15'978
DISTEMIST-ES (Miranda-Escalada et al. 2022)	8'065	5'136
DISTEMIST-FR (Miranda-Escalada et al. 2022)	6'447	6'445
DISTEMIST-FR (ours)	8'059	5'132

Entity Linking Quality

Top 10 concepts order **retained** after the projection.

Most frequent annotations per concept are **consistent**.

1. Miranda-Escalada, A., Farré, E., & Krallinger, M. (2020). Named Entity Recognition, Concept Normalization and Clinical Coding: Overview of the Cantemist Track for Cancer Text Mining in Spanish, Corpus, Guidelines, Methods and Results. IberLEF@ SEPLN, 303-323.
2. Miranda-Escalada, A., Gascó, L., Lima-López, S., Farré-Maduell, E., Estrada, D., Nentidis, A., ... & Krallinger, M. (2022, September). Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. In CLEF (Working Notes) (pp. 179-203).

SIMED@HUG SIMED@UNIGE

UNIVERSITÉ DE GENÈVE FACULTÉ DE MÉDECINE

Hôpitaux Universitaires Genève

SIMED