BIOLOGICAL TEXT MINING UNIT

Martin Krallinger
Head of Unit

Staff Scientists
José Antonio López (until January),
Marta Villegas

OVERVIEW

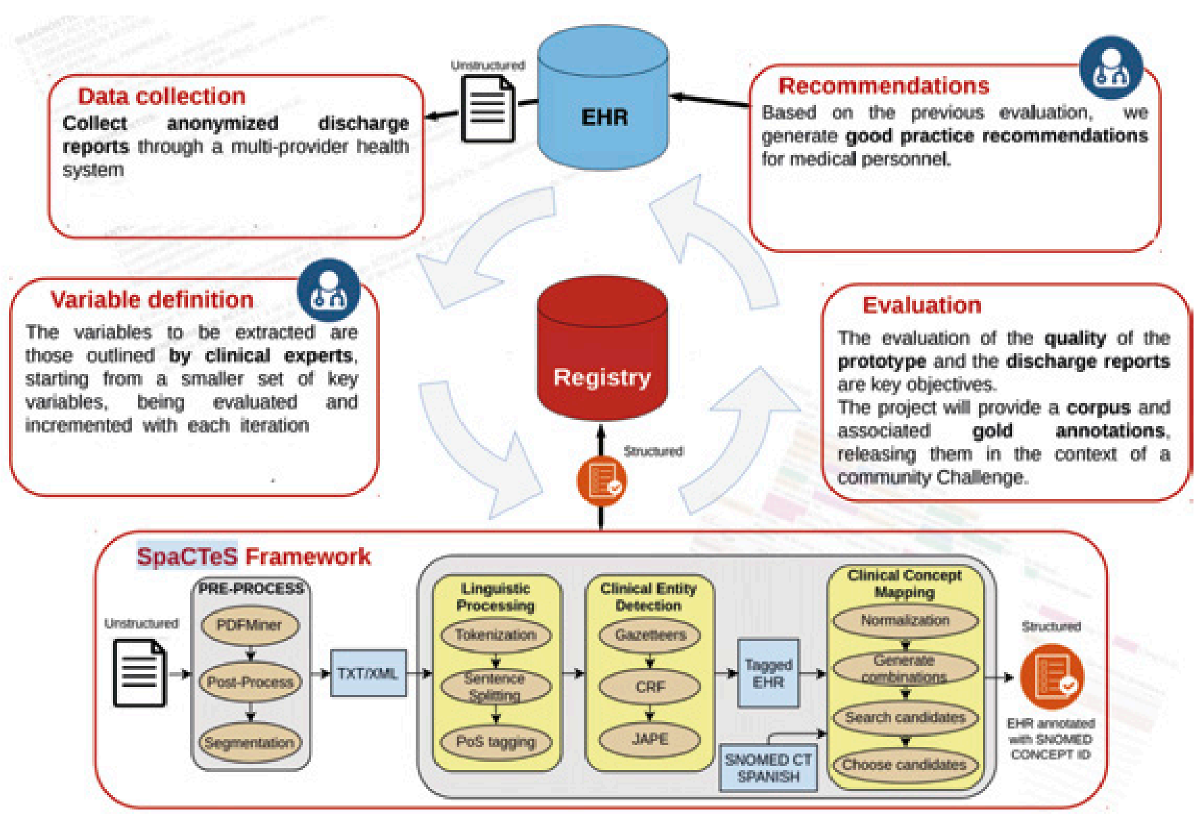
Biomedical cancer research is a particularly data-heavy discipline, where key information sources are not only limited to genomic information or raw experimental data. Especially unstructured data, such as the scientific literature, clinical texts, medicinal chemistry patents or patient generated content, constitute a valuable resource for a range of scenarios like drug discovery, interpretation of large scale experimental results, drug repurposing or evidence based medicine. Medical big data approaches are only able to efficiently exploit running texts through the use of natural language processing (NLP) techniques relying on deep learning and artificial intelligence strategies. Our Unit is financed through the Plan for the Advancement of Language Technologies; the aim is to generate resources that can improve the exploitation of biomedical data by means of implementing and evaluating the underlying quality of systems for automatic recognition of medical concepts, generation of specialised neural machine translation models for the medical domain and the implementation of a medical language technology platform and software components for processing Spanish EHRs.

> "Language technologies, together with artificial intelligence, are driving the technological transformation of biomedical and clinical data into actionable information at all levels of cancer research."

Technicians
Aitor González (TS)*, Ander Intxaurrondo (TS)*, Jesús Santamaría (TS)*

*Titulado Superior (Advanced Degree)

RESEARCH HIGHLIGHTS



Collect anonymized discharge reports through a multi-provider health system

Based on the previous evaluation, we generate good practice recommendations for medical personnel.

The variables to be extracted are those outlined by clinical experts, starting from a smaller set of key variables, being evaluated and incremented with each iteration

The evaluation of the quality of the prototype and the discharge reports are key objectives.
The project will provide a corpus and associated gold annotations, releasing them in the context of a community Challenge.

Figure Clinical NLP framework for processing electronic health records in Spanish and Catalan.

The Biological Text Mining Unit has provided consultancy, guidance and technical support for clinical text mining use cases posed by several healthcare institutions (*Hospital Virgen del Rocio, Hospital XII de Octubre, Hospital Son Espases, Hospital Clinic)*, national and regional health-related agencies (Spanish Medical Agency, *Instituto Aragonés de Ciencias de la Salud, Servicio Andaluz de Salud, Fundació TIC Salut Social*), and natural language as well as medical informatics academic research groups. The Unit has contributed to benchmarking efforts of clinical text mining systems by organising shared tasks in the context of community challenges organised by the *Sociedad Española para el Procesamiento del Lenguaje Natural* (*SEPLN-IberEval*) and releasing high quality evaluation datasets. The Unit has published a collection of clinical NLP resources, all freely available at: https://zenodo.org/communities/medicalnlp and https://github.com/PlanTL. In addition to annotation guidelines and Gold Standard corpora for developing and evaluating the quality of systems for automatically detecting biomedical and clinical concepts, the Unit has implemented software tools for automatic medical term recognition and normalisation (CUTEXT), an electronic health record sectionizer, a medical sentence boundary recognition system, a medical text tokenizer, lemmatizer and PoS-tagger. Moreover, we have also contributed to the first Protected Health Information (PHI) masker for the Spanish language, a system for medical negation detection, clinical temporal expression detection based on HeidelTime, a medical machine translation system and word embeddings. These key constituents are being integrated into the clinical NLP pipeline developed by the Unit. ■