Text mining

Biomedical and clinical research is a particular data heavy discipline, where key information sources are not only represented by 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 re-purposing or evidence based medicine. Medical big data approaches are only able to exploit efficiently 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 with the aim of generating 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 specialized neural machine translation models for the medical domain and the implementation of a medical language technology platform and software components for processing Spanish EHRs.

The Text Mining Unit focuses on the application and development of biomedical text mining technologies, which are becoming a key tool for the efficient exploitation of information, contained in unstructured data repositories including the scientific literature, electronic health records (EHRs), patents, biobank metadata, clinical trials and social media. The unit has a particular interest in processing clinical documents written in Spanish and other co-official languages in the area of health-related topics and the integration of molecular and biological information derived from the literature.

The 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 organizing shared tasks in the context of community challenges organized 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 normalization (CUTEXT), a electronic health record sectionizer, a medical sentence boundary recognition system, a medical text tokenizer, lemmatizer and PoS-tagger. Moreover we have also contributed the first Protected Health Information (PHI) masker for Spanish, 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.

Objectives

The strategic goals of the Text Mining Unit are to:

- Design and develop biomedical language-processing resources with emphasis on oncology.
- Provide consultancy and technical advice for language technologies in the biomedical domain.

- Design requirements and standards for interoperability of biomedical language technologies.
- Coordinate community assessment and evaluation challenges of biomedical text mining tasks.
- Leveraging the uptake of biomedical text mining technologies and relevant standards.

One of the main scopes of the unit is to provide biomedical text mining and language processing infrastructures that can be maintained efficiently over time and be integrated in biomedical analysis platforms comprising data from experimental outcomes of patient-derived information.

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---