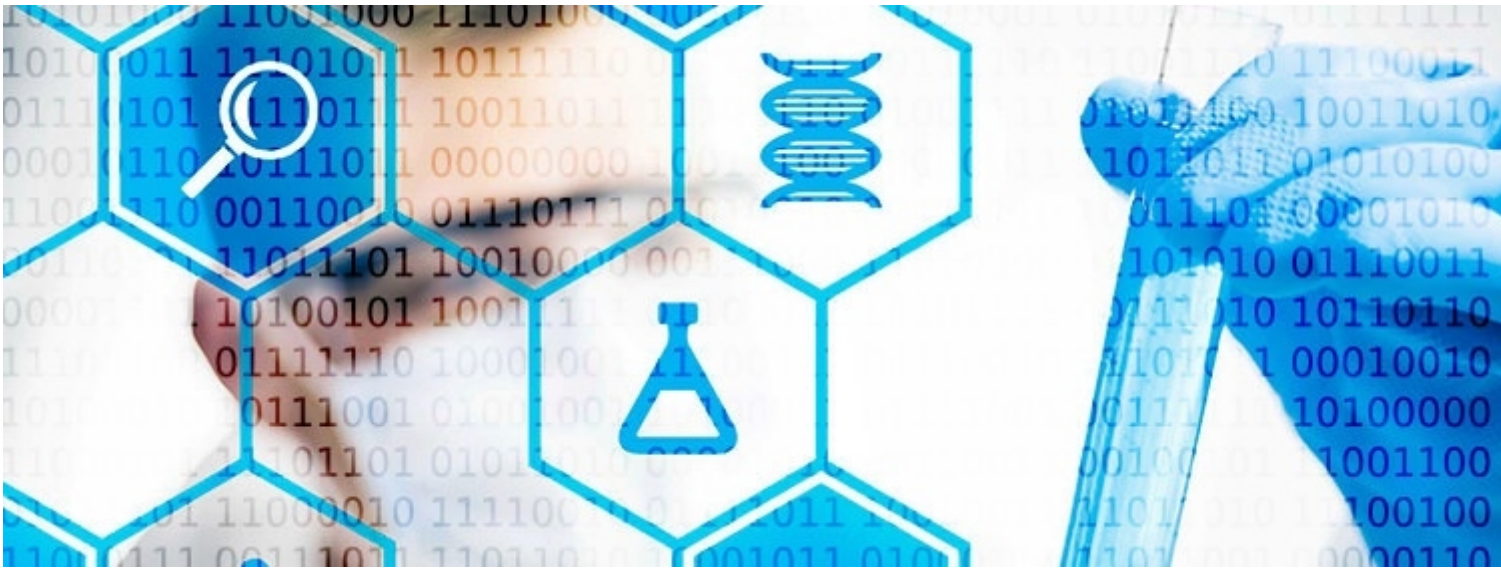
Spanish researchers review the state-of-the-art text mining technologies for chemistry

A considerable fraction of cancer-relevant data is only available in the form of unstructured biomedical healthcare data.



In a recent *Chemical Reviews* article, the Biological Text Mining Unit at the Spanish National Cancer Research Centre (CNIO) in coordination with researchers at the Center for Applied Medical Research (CIMA), at the University of Navarra, and at **Barcelona Supercomputing Centre (BSC)** have published the first exhaustive revision of the state-of-the-art methodologies underlying chemical search engines, named entity recognition and text mining systems. **Alfonso Valencia**, former researcher at CNIO and currently at BSC as Life Sciences Department Director, was the senior researcher who supervised the work. BSC is particularly interested in exploring and developing the application of HPC and Machine Learning techniques to the extraction of information in areas of chemistry and biomedicine.

The rapidly growing field of big data applications in biomedical research together with the use of machine learning and artificial intelligence technologies for text data mining has resulted in promising tools. “This review –state the authors– is organized to serve as a practical guide to researchers entering in this field but also to help them to envision the next steps in this emerging data science field”.

“Through the release of Gold Standard datasets and the organization of several community challenge benchmark events, the Biological Text Mining Unit has played a critical role in the development and evaluation of current chemical text mining systems, as highlighted in this article,” explains Martin Krallinger, head of the Unit and first author of the review.

A huge amount of unstructured data

A considerable fraction of cancer-relevant data is only available in the form of unstructured biomedical healthcare data. This type of data includes the rapidly growing scientific literature, medicinal chemistry patents, electronic health records or clinical trial documents. In fact, every year, over 20,000 new compounds are published in medicinal and biological chemistry journals.

Being able to transform unstructured cancer research data into structured databases that can be more efficiently processed by machines or queried by humans is becoming critical for a range of very heterogeneous applications. These include the identification of new drug targets, re-purposing of approved drugs, the identification of adverse drug events or retrieval of systems biology associated with chemical-disease or chemical-gene networks. Chemical compounds and drugs constitute a key entity type of particular relevance for biomedical research.

This review provides comprehensive and in-depth description of fundamental concepts, technical implementations, and current technologies for meeting these information demands.

Reference:

Krallinger M, Rabal O, Lourenço A, Oyarzabal, Valencia A (2017). Information Retrieval and Text Mining Technologies for Chemistry. Chemical Reviews

Nota en castellano (pdf)
Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---