Mining Biomedical Text to Find Insight that Can Save Lives

Blog post by Davide Cirillo, Alejandro Canosa, and Salvador Capella-Gutierrez (Barcelona Supercomputing Center), Maria Rodriguez Martinez and Matteo Manica (IBM), Joao Pita Costa and Jolanda Modic (XLAB)

The exploitation of knowledge accumulated in the increasing amount of published research represents a challenging task for the scientific community but, at the same time, it allows us to identify new opportunities and key ideas. In this regard, the advances on Natural Language Processing (NLP) have led to the optimization of this common knowledge, helping researchers in analysing and assessing issues and problems specific to areas of action, such as cancer research. This is particularly important when exploring pediatric cancers due to the lack of data when these cancers are less common. Text mining technologies and techniques can help researchers find insightful information within text-based data sets that cannot be analysed manually or with the existing approaches for structured data types. NLP enables to identify new procedures, best practices and success stories, as well as to establish data exploration tools based on free text and consider ontological constructs to better utilise and potentiate the value of that data. Moreover, open data initiatives led by the scientific community and government institutions, make it possible to consider meaningful and effective NLP approaches that support Healthcare activities.

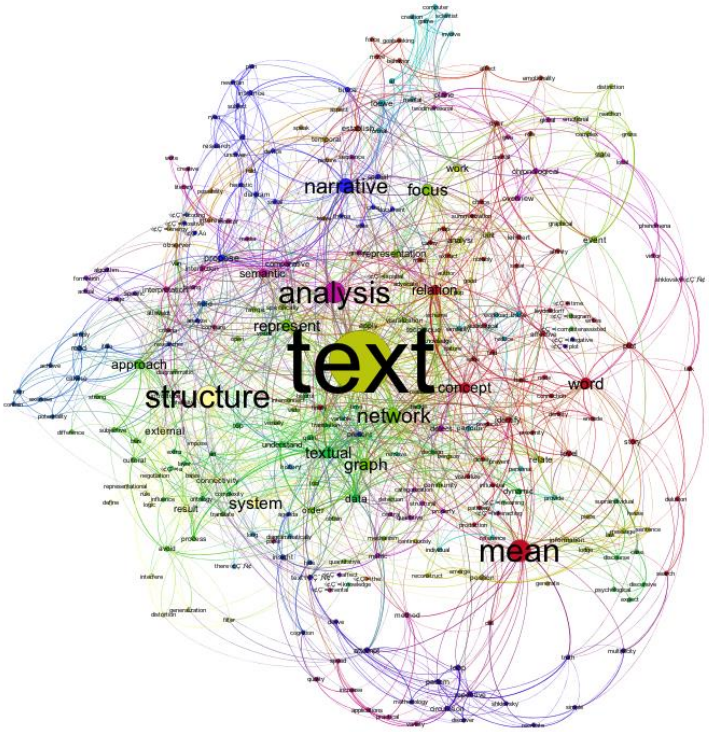

Figure 1 - An illustration of the text mining capabilities, extracting information from text, using machine learning algorithms to capture relations between entities that can be used in cancer research

Nowadays, there is a wide range of available technologies and data sources that can provide relevant contributions to Healthcare professionals. With most research results being available on the internet, it is now possible to access ideas, concepts and methods that would be almost unreachable a few years ago. Texts written in natural languages form a big part of the available data that can be used by healthcare professionals to drive innovation in knowledge domains, for instance, in cancer research where the multitude of approaches is overwhelming, ranging from purely genetic-based methods [3] to topological data analysis [14]. NLP methods are usually developed as general-purpose tools, but can nevertheless achieve good performance for many application domains. However, in order to increase the quality of results, it is often necessary to fine-tune such tools for a specific application, for instance, using specialized training sets, redefining keyword definitions, category definitions or, in general terms, refining the model behind a specific algorithm. The nature of the datasets varies within the targeted research problems.

With the advent of widespread data analytics and visualisation positively affecting the daily life of people worldwide, biomedical researchers are leveraging the facilitated access to efficient technological tools to better profit from that advantage. Particular examples of this are PubMed, a central reference to state-of-the-art medical research, and its European counterpart EuropePMC, an ELIXIR Core Resource that supports the continuous integration of new text-mining applications for the advanced exploitation of the scientific literature. This search engine tool is frequently used to have an overview of a certain topic using several filters, tags and advanced search options. It has been freely available since June 1997, providing access to references and abstracts on life sciences and biomedical topics. MEDLINE is the underlying open database serving the PubMed engine, maintained by the United States National Library of Medicine (NLM) at the National Institutes of Health. It includes citations from more than 5,200 worldwide journals in about 40 languages (about 60 languages in older journals). In 2020, PubMed has more than 30 million records dating from 1946 to the present day. About 500,000 new records are added each year. 17.2 million of PubMed's records are listed with their abstracts, and 16.9 million articles have links to full-text, of which 5.9 million articles have full-text available for free online. In particular, it includes 443,218 full-text articles with the keywords string “public health”.

There have been several initiatives to mine the MEDLINE open dataset, including biomedical natural language processing approaches (BioNLP) [2], which focused on biomedical text mining combining controlled vocabularies and ontologies available through the National Library of Medicine's Unified Medical Language System (UMLS) [4] and Medical Subject Headings (MeSH) thesauri; the semantic biomedical question answering system SemBioNLQA for retrieving information based on natural language questions [15]; the usage of topic models to enrich the meta-information provided by MeSH [13]; and the foundational work on the medical indexing expert system, MedIndEx, using knowledge-base frames to guide indexers in completing indexing frames [7]. More general approaches include the Medical Text Indexer (MTI) [1], that provides MeSH indexing recommendations to support the human indexers of the NLM using k-nearest neighbors (KNN), pattern matching and indexing rules; and the MeSH Now [12], including a learning-to-rank framework achieving good results in its automated classification. The NLM also made available other useful tools like, e.g., the MeSH on Demand [16], that suggests MeSH vocabulary explicitly mentioned in the input text; and the Semantic MEDLINE [10], that aims for the semantic knowledge representation of MEDLINE itself.

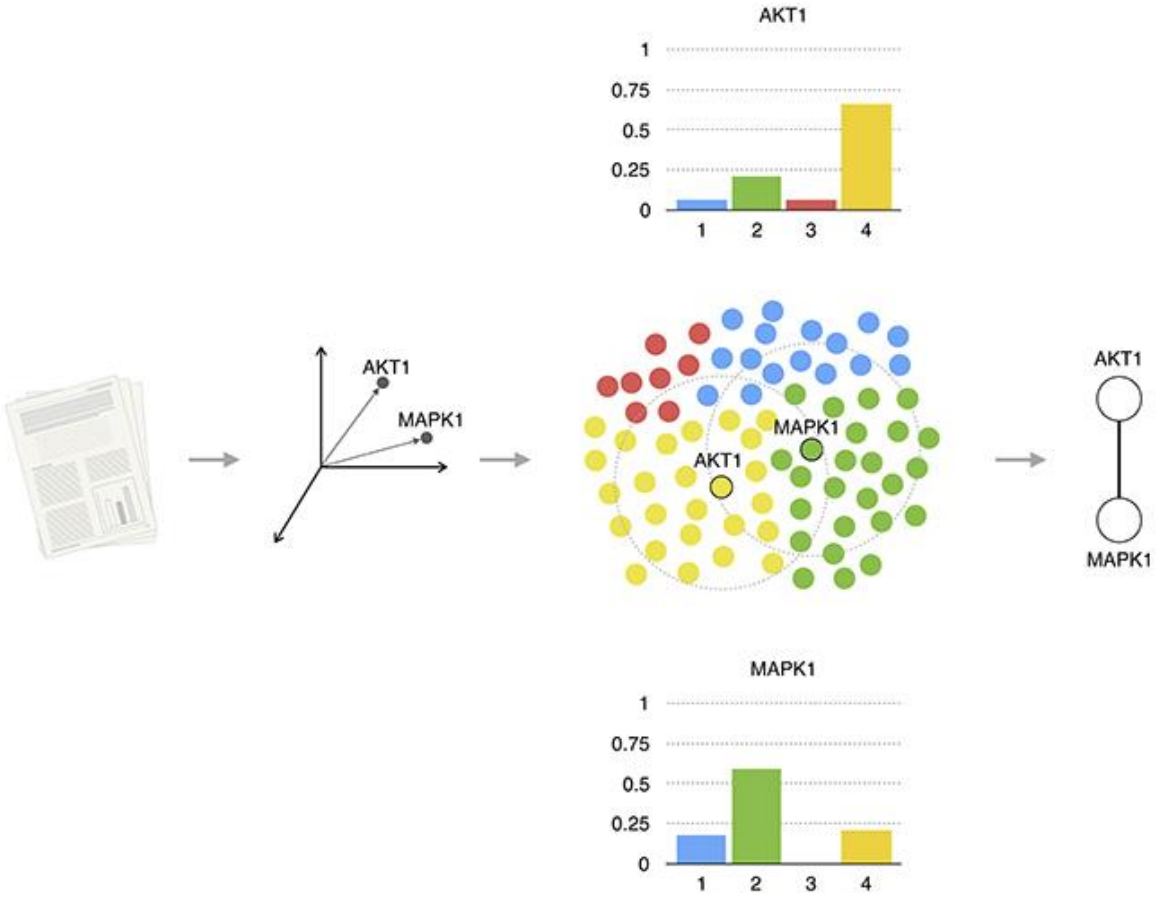

Figure 2: INtERAcT enables mining of unstructured text data in scientific publications, to extract valuable information on relations between entities of interest, e.g., protein-protein interactions, automatically and with no requirements for human annotation [11].

The advanced NLP methods have also been proven useful in the context of cancer research, particularly when focusing on specific aspects of research. Examples of these include: INtERAcT [11], automating knowledge extraction on protein-protein interactions from scientific research in a completely unsupervised way; MelanomaMine [6], a text mining application and database dedicated to the processing of melanoma-related biomedical literature and knowledge resources; LimTox [5], a tool for extracting associations between compounds and a particular toxicological endpoint from the content of toxicology reports at various levels of granularity and evidence types. All these methods are intended to serve basic researchers and medical personnel as a domain-specific search and information retrieval systems for oncology. For instance, LimTox facilitates the establishment of toxicity thresholds of several substances, while MelanomaMine allows the detection of bio-entities of relevance (e.g., genes, proteins, mutations and chemicals/drugs) to understand the molecular basis of melanoma.

Related to this topic, we also highlight the initiative OpenMinTeD [9], enabling an infrastructure that fosters and facilitates the use of text mining technologies in the scientific publications world, builds on existing text mining tools and platforms, and renders them discoverable and interoperable through appropriate registries and a standards-based interoperability layer, respectively. It supports the training of text mining users and developers alike and demonstrates the merits of the approach through several use cases identified by scholars and experts from different scientific areas, ranging from generic scholarly communication to literature related to life sciences, including cancer research. Terminology and language expressions in cancer research make the analysis of biomedical unstructured data a notoriously difficult domain for NLP and text mining. In these areas, relevant sources comprise heterogeneous scientific literature in various highly specialized subdomains with a highly ambiguous language, characterized by acronyms, abbreviations and ever-changing technical terms. Initiatives such as OpenMinTeD as well as the bi-annual challenge BioCreative that the Life Department at BSC organizes [8] are contributing to substantial progress in NLP and the extraction of bio-entity mentions and their relations.

In the iPC project, we are making use of these NLP knowledge extraction methods and initiatives to progress on pediatric cancer research. We are implementing an NLP system for mining large volumes of paediatric cancer-related abstracts in NCBI PubMed. The system will be built on LiMTox and MelanomaMine approaches originally developed for melanoma and hepatotoxicity but applied to paediatric tumours. OpenMinTeD and BioCreative community-wide efforts will provide evaluation criteria and guidelines to curate, validate and compare this text workflow in terms of interoperability and performance.

The system will also leverage INtERAcT functionalities that can be used to extract relations, such as protein-protein interactions, exploiting word embeddings. This technology for language modeling based on deep learning does not require text labeling for training or domain-specific knowledge, and hence can be easily applied to different scientific domains in a completely unsupervised way. Here we will adapt and generalize INtERAcT in three different directions. First, we will generalize the search technology in the embedding space to extract disease-specific molecular interactions (e.g., drug-target interactions in childhood leukemia) without having to build a new embedding for each paediatric tumour. Second, we will integrate the results generated by LiMTox and MelanomaMine in INtERAcT to extract other types of relevant two-entities interactions, such as drug-target, mutation-disease, drug-disease interactions, etc. Finally, we will expand INtERAcT to predict tertiary interactions, e.g. drug-gene-disease, that will be important in downstream WPs to implement the virtual patient personalized models.

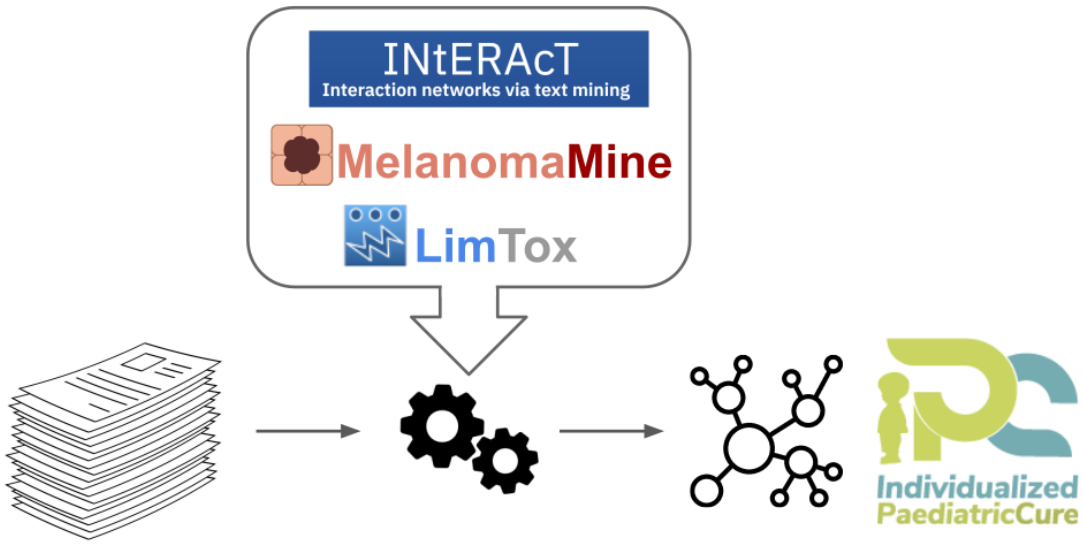

Figure: The usage of text mining in iPC

Our implementations will generate biomedical knowledge to be utilized by the iPC consortium facilitating tasks such as the identification of cross-tumour similarities, the modelling of drug modes of action and patient susceptibility to specific drugs, the detection of public data including metadata associated with paediatric cancers, among others.

We plan to continuously refine our NLP approaches to better characterize the identified bio-entities and relationships and derive actionable insights by integrating this information with the multi-omics data available to the iPC consortium. This integration will be instrumental to the identification and characterization of biomarkers and relations that can be, in turn, utilized by the modeling components of the project.

The European Genome-phenome Archive (EGA) is one the largest European platforms for sharing and reusing personally identifiable genetic and phenotypic data resulting from biomedical research projects. Although metadata querying is the most common way to retrieve EGA datasets, the inherent fuzziness of categorical descriptors, as well as the presence of synonyms and similar concepts, make this practice extremely inefficient. We will apply the text mining technologies integrating LiMTox, MelanomaMine and INtERAcT to develop an efficient query system for the EGA metadata and associated scientific publications. We will work closely with the EGA team to improve metadata querying and enhance domain-specific knowledge extraction at the time that exploits the knowledge existing in publications associated withdatasets deposited in EGA

[1] A. Aronson et al (2004). The NLM indexing initiative's medical text indexer. Medinfo, vol. 89.

[2] S. Baker and A.L. Korhonen (2017). Initializing neural networks for hierarchical multi-label text classification. Association for Computational Linguistics. BioNLP 2017, Association for Computational Linguistics. pp. 307–315.
[3] N. J. Birkbak, and N. McGranahan. "Cancer genome evolutionary trajectories in metastasis." Cancer Cell 37, no. 1 (2020): 8-19.
[4] O. Bodenreider (2004). The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl_1), D267-D270.
[5] Cañada, Andres, et al. "LimTox: a web tool for applied text mining of adverse event and toxicity associations of compounds, drugs and genes." Nucleic acids research 45.W1 (2017): W484-W489., http://limtox.bioinfo.cnio.es/
[6] Centro Nacional de Investigaciones Oncológicas (CNIO), http://melanomamine.bioinfo.cnio.es/
[7] S. M. Humphrey (1989). "MedIndEx system: medical indexing expert system." Information Processing & Management 25.1 (1989): 73-88.
[8] IBM, http://www.biocrative.org
[9] IBM, http://openminted.eu/
[10] H. Kilicoglu et al (2008). Semantic MEDLINE: a web application for managing the results of PubMed Searches. In Proceedings of the third international symposium for semantic mining in biomedicine. Vol. 2008, pp. 69-76.
[11] M. Manica, R. Mathis, J. Cadow, and M. Rodríguez Martínez. "Context-specific interaction networks from vector representation of words." Nature Machine Intelligence 1, no. 4 (2019): 181-190.
[12] Y. Mao and L. Zhiyong (2017) "MeSH Now: automatic MeSH indexing at PubMed scale via learning to rank." Journal of biomedical semantics 8.1: 15.
[13] D. Newman, S. Karimi, and L. Cavedon (2009). Using topic models to interpret MEDLINE's medical subject headings. In Australasian Joint Conference on Artificial Intelligence, pp. 270-279. Springer, Berlin, Heidelberg, 2009.
[14] R. Rabadán et al. (2020). Identification of relevant genetic alterations in cancer using topological data analysis. Nature communications, 11(1), 1-10.
[15] M. Sarrouti and Said Ouatik El Alaoui(2020). SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions. Artificial Intelligence in Medicine 102 (2020): 101767.
[16] P. Srinivasan and B. Libbus (2004). Mining MEDLINE for implicit links between dietary substances and diseases. Bioinformatics, 20(Suppl 1), i290–i296.