Revealing disease similarities by text mining

Alberto Calderone[1], Luana Licata[1], Elisa Micarelli[1], Livia Perfetto[1], Gianni Cesareni[1]

[1] Bioinformatics and Computational Biology Unit, Department of Biology, University of Rome 'Tor Vergata', Rome, 00133, Italy

Abstract. Texts written in human language contain structured information that is not easily parsable by computers. Text mining relies on large text corpora to derive rules which can be used by automatic means to extract automatically such information.
Scientific literature represents the main source of information to study any biological phenomenon. While some phenomenon are studied to the point that corpora can actually be build, scientific literature describing rare diseases is scarce implying an even bigger challenge for automatic approaches.
In order to tackle this problem the ELIXIR infrastructure is supporting various initiatives for data integration in different field of life sciences, including rare diseases, which will pave the way to the development of dedicated pieces of software.
In this work we present a tool which applies a text-mining strategy to multiple text sets and merges individual results in order to infer not explicitly written connections.

Keywords: Text mining, rare diseases, data integration.

1 Overview

In order to get the current understanding of a research topic a scientist needs to read through scientific literature. Studying many articles is the key to making mental connections and come up with hypotheses that are not yet explicitly reported. This mental process is far from being a trivial task even for experts. Automatic text analyses can support and facilitate information extraction by speeding up tasks such as keywords identification.

Text mining (TM) uses statistical and computational approaches to derive text patterns which can in turn extract useful information. Some TM tools aim at highlighting keywords or phrases usually relying on training corpora. While some topics are studied to the point that corpora can actually be build, scientific literature describing rare diseases is scarce implying an even bigger challenge for automatic means.

ELIXIR is an European initiative which sustains bioinformatics resources across member states. ELIXIR aims at making Europe's science institutes and organizations come together under the same hood to manage the increasing amount of data being generated in the field of life science research. It supports various initiatives for data integration and dissemination which will pave the way to the development of dedicated pieces of software. These data can be used to instruct machine learning systems in

conjunction with text mining tools in order to extract information from this scarcely explored field of life sciences.

While keywords identification in texts is useful when reading articles to spot important words and phrases, much information is usually scattered in many articles and possibly in articles from different domains or topics which can often only be derived by mental reasoning. To support mental reasoning and linking of idea, we developed a tool which aims at analyzing and integrating multiple articles in order to extract not explicitly written information.

Publication abstracts and titles mentioning a disease are retrieved from the structured data returned by PubMed. These documents are preprocessed to remove unnecessary terms, lemmatized and tokenized.

There exist several TM approaches to extract keywords from text. For instance, term-frequency, term frequency–inverse document frequency [1], Parts-Of-Speech (POS) tagging algorithms [2], and others. We are currently investigating the best approach to extract gene names from scientific literature. At the moment, we are relying on a method based on statistics using exact word matching.

Relevant terms are ranked according to a p-value calculated against a random set of articles and then compared versus a second query results, for instance, a second disease. The most relevant terms identified are represented as vectors of real values whose distances can be calculated as a representation of their semantic distance. Using these distances we build a graph which links diseases according to their similarities (see Fig. 1).

In particular, we applied two approaches, one based on MeSH terms [3] vocabulary and one on full text analysis. As a preliminary analysis we processed articles about diseases in general, including rare diseases. This preliminary analysis allowed us to cluster diseases according to similarities. In particular, rare diseases turn out to be associated with other well studied conditions triggering possible connections.

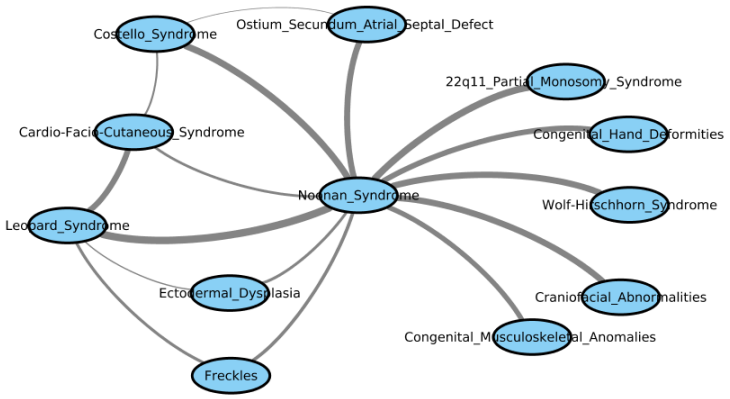

Fig. 1. A small subgraph of the disease similarity graph built from literature using symptoms extracted from MeSH terms. Noonan, Costello and Leopard syndromes are connected as expected. In particular, we reported the top-five diseases similar to Noonan Syndrome.

2 Future Developments

Recently we published a database focused on diseases DISNOR [4] . Since one of our main goals is to increase the coverage on rare diseases, we are planning to integrate TM in our data duration pipeline.
We plan to improve our TM software whose results will also benefit from two ELIXIR case studies.

1) The dedicated APIs which are being developed by the ELIXIR nodes involved in the RD-connect [5] will contribute to set-up a common interface to different data sources about rare diseases.
2) The integration of various data sources will pave the way to the development of dedicated TM tools to extract information about rare diseases and eventually to a more precise terms and concepts extraction.

References

1. Rajaraman, A., Ullman, J.D.: Mining of Massive Datasets. Cambridge University Press, Cambridge (2011).
2. Brill, E., Eric: A simple rule-based part of speech tagger. In: Proceedings of the third conference on Applied natural language processing -. p. 152. Association for Computational Linguistics, Morristown, NJ, USA (1992).
3. ROGERS, F.B.: Medical subject headings. Bull. Med. Libr. Assoc. 51, 114–6 (1963).
4. Lo Surdo, P., Calderone, A., Iannuccelli, M., Licata, L., Peluso, D., Castagnoli, L., Cesareni, G., Perfetto, L.: DISNOR: a disease network open resource. Nucleic Acids Res. (2017).
5. Thompson, R., Johnston, L., Taruscio, D., Monaco, L., Béroud, C., Gut, I.G., Hansson, M.G., 't Hoen, P.-B.A., Patrinos, G.P., Dawkins, H., Ensini, M., Zatloukal, K., Koubi, D., Heslop, E., Paschall, J.E., Posada, M., Robinson, P.N., Bushby, K., Lochmüller, H.: RD-Connect: An Integrated Platform Connecting Databases, Registries, Biobanks and Clinical Bioinformatics for Rare Disease Research. J. Gen. Intern. Med. 29, 780–787 (2014).