Text mining and biological literature

Science, Genetics

TEXT MINING AND BIOLOGICAL LITERATURE

Introduction

We live in the era of digital age, the era where knowledge and information is power and considered to be the key in our overall progress. Every day, hundreds of documents appear in digital format in the the internet in the form of articles in newspapers or magazines, online books, scientific discoveries and publications. Without doubt, we can say that knowledge is infinite, the amount of information we have in our hands is endless. In the fields of biology, biotechnology, medicine and their subfields, every week thousands of articles are being published online regarding established reviews and analytics, discoveries and new experiments. This perpetual river of information should be extractable by scientists in order to continue their researches, update their work and finish their experiments. All this biomedical data is growing day by day and we can say that handling this infinite data is dire in the aforementioned research fields. Growing number of literature has become a significant problem for scientist and researchers. This sophisticated amount of literature is becoming impossible to follow even for the most experienced readers and it can lead to a waste of money and research time.

What is genomics?

A genome represents the entire DNA substance that is present in one cell. Using DNA sequence techniques and bioinformatics we can analyze the structure and function of a genome. We can study how genes interact with each other or/and with the environment they reside. The scientific field

focusing on the operation of genomes is called Genomics. Genomics experts attempt to unfold all the mysteries of the DNA sequence in order to give answers to complex challenges. For example, genomics focuses on the examination of genomes appearing in serious diseases such as cancer, diabetes, heart diseases and many more. As you can imagine, a huge amounts of new data is being developed every day in this research area. And yes, keeping up with this kind of information is not easy.

A new field : Text mining

As we mentioned before, in the numerous subfields of biology and medicine we get tons of information in the form of numbers, sequences and genomes but we also get something else. As a logical consequence, tons of plain text join the scientific publications. It is the essential “ literature”, where scientists describe their thoughts, explain their methodology and analyze their conclusions. This textual data is considered to be a great tool for those who can handle it and use it accordingly. At the same time though, this phenomenon generates a question and creates a new field of research. The arising question is the following: “ Are we able to handle this amount of data ? And if yes, how will we find whatever we are looking for in time in order to make the right decisions?

This new research field is called Text Mining. TM , as we are going to call it from now on, focuses exclusively on discovering and extracting unknown literature texts by combining sophisticated methods of machine learning, computational linguistics and informational retrieval. By using these techniques we will be able to gain significant time in information extraction,

which will lead to a more promising hypothesis generation. In human genomics this automated gene and protein detection seems very promising, as we have an significant new amount of reports establishing new variables about rare diseases. Being able to study, evaluate and connect this new variables to existing information is crucial. We have to mention though that due to copyright reasons very few articles are free to read, comparing to the vast ocean of publications online, hence TM is focusing on titles and abstracts which are freely accessible in databases such as BMC and MEDLINE.

How TM works?

TM, as we mentioned earlier, is about discovering unstructured knowledge. Most of times we have to deal with three major objectives: identifying the essential data, extracting the information and detecting the associations between the already extracted data. We can imagine TM as a curator, who searches all the available resources such as online publications, patents, journals and so on, finds the available texts, links them all together and categorizing them. First of all, if we want to extract biological literature from a text we should be able to identify it. Biological entities can be proteins, cells, genes, genomes, diseases, chemical compounds and many more others biological definitions. Then we will have to do Named entity recognition (NER) and Terrn Normalization that is, distinguishing, storing and sorting into categories our findings and associating them with the right entities in our database. Next step will be to check the relation between the stored entities, define what kind of relation that is as well as the type of it. But lets talk a bit more about this stages: NEROur first thoughts considering

NER should focus on two problems: First of all, the ever-evolving literature of biology. There are millions of definitions referring to genes, proteins, patterns, compounds etc and many more are being created as we write this very text. Secondly the similarity in acronyms or abbreviations in biological terminology and the variety of definitions an entity can have. For example, entities P53, TP53 and TRP53 relate to the same gene or when our imaginary curator come across with the word “ Parkinson” it has to make a choice and decide whether it is referring to James Parkinson who was first to study Parkinson’s disease or the disease itself?

In order to address these problems a new committee was created. The HUGO Gene Nomenclature Committee (HGNC) targets on appointing a unique name and symbol for all known genes and until now HGNC has done a great job assigning names and symbols in over 35. 000 entities. This number is big but there are more entities out there still unassigned.

There are three main methods used for NER (Hybrid methods can also be used):

- Dictionary-based
- Rule-based
- Machine learning

Dictionary-based: These methods uses a simple text-matching algorithms with a preset dictionary. We search text and then we match our findings with the entities of our dictionary. Dictionary-based techniques are extremely dependent on our preset dictionaries and the matching algorithms we use and that is why they develop a large number of ambiguous results. Ruled-

based: This kind of methods focuses on recognizing entities based on symbols, numbers, and suffixes/affixes. For example, many biological entities end with specific suffixes such as -in such as, Keratin – A fibrous structural family or Myosin – Motor proteins known for taking part in muscle contraction, etc. So, this methods create rules that helps them categorize words with specific orthographic features as teams. Ruled-based methods are considered very accurate as with a simple rule they can classify a big number of entities.

On the other hand, due to the variant grammatical and syntactical rules of our language, they are not so agile. Machine learning: Machine learning methods are considered to produce the best results for NER. They use large amounts of annotated data sets in order to identify and classify entities of text. We have two major machine learning techniques: Classification and Sequence labeling. Today, these methods are being used increasingly compared to ruled-based and dictionary-based methods. Term normalization Now that we have found and flagged out results by using NER methods, we must link them with the appropriate entries in our databases. Term normalization compare entities and assign the matching identifier. Today, one of the most used knowledge databases is the Gene Ontology (GE) which aims in the development of a computational model describing the properties and functions of genes. Here we have to mention again the difficulty in associating and matching entities based on the biological literature. The genomic nomeclature is rich and at the same time ambiguous (genes/proteins can result to more than one identifiers).