Text Mining in Bioinformatics

Arda Çelebi
Department of Computer Engineering
Boğaziçi University
İstanbul, Turkey

1 Introduction

Every other day, more and more we come to realize that knowledge and information becomes the driver of the progress from business to science. Every deal businessman makes, every experiment scientist conducts must rely on some sort of information that justifies that taken decision. In the era of the digital age, considering that we not only publish every bit of information in digital format but also let people add their thoughts and experiences onto those published content in such a social environment of the Internet, the amount of knowledge we can access is virtually limitless. However, even though the knowledge is infinite, we have to make our decisions in finite time, and fast. So the problem is finding the right information in the digital ocean of knowledge, which is pretty bigger endeavor than finding the needle in the haystack.

The real workers of the digital age are not the humans but the computers. However so far best they can do is to obey our well-defined commands. It is proven that making a computer understand human language is a very difficult task, or as some consider impossible. Nevertheless, people who have been working on that subject for years have developed techniques that make computers detect, distinguish, and extract certain types of information. Even though computers couldn't understand the reality expressed between the lines, they can somewhat mimic our way of doing things with words. To give an example, they can replace the French words with their equivalent English words and then reorder the resulting sequence of words to come up with the translation of that given French sentence, all in automated and rule-based fashion with no understanding involved. The outcome may not be the perfect translation but can be considered as enough to understand what the original sentence is saying.

Aforementioned example, which is from the field of Machine Translation (MT), is a perfect case showing that computers may not need to understand the sentence, even though they have to in order to get the exact translation. But, after all, not every human translator is perfect, too. Compared to MT task, making the computer finding specific information is relatively simpler task. Just consider the search engines, such as Google. Everyday we search among zillions of documents on the Internet and most of the time we get satisfied with returned documents. As in the case of MT, searching documents, or what it is called Information Retrieval (IR), does not require computer to understand the content of any document. However, like in the example of the exact translation, if we are looking for exact answer to our query, computer has to be able to locate asked information in those searched documents. Apart from matching the query with documents, this also involves extraction of knowledge conveyed in the document so that computer can present that extracted information as the exact answer for given query. All detection, extraction and retrieval of the knowledge is the subject of Text Mining (TM).

Today, two of the biggest rivers that feed the digital ocean of knowledge are the fields of biology and medicine, not to mention their countless subfields. Every day, while some scientists look for specific knowledge that support their thought processes and experiments, some others finish their experiments and publish their findings. In this cycle, every useful bit of information must be mined out of that published content so that scientists who are looking for a particular information can get
their exact answers, if there is any. This case study investigates this endeavor, that is mining the infinite stream of text in the context of bioinformatics.

2 Biology, Biomedical Research, and Bioinformatics

While in the field of biology, scientists work on the biological processes happening in a body, scientists in the field of medicine conduct biomedical research to understand how medicine interacts with those biological processes and to put that knowledge into practical use in the forms of diagnosis, prevention, and treatment. For both fields, knowledge is the driver of the progress. While it is all in digital, this information can be in any format. It can be sequences of characters representing the part of the DNA, bunch of numbers coming out of an experiment or published text in natural language.

Apart from data in the form of sequences of proteins and numbers, the real treasure lies within the lines of textual data, that is published findings of scientists. However, the fact that every day more and more papers are published by the scientists makes finding the right and most recent information in this growing mountain of knowledge virtually impossible. To exemplify how serious this growth is, we can look at the expansion of the MEDLINE database, which maintains the abstracts of papers in the field of biomedical research. In 2004, the statistics showed that 500,000 new papers were included to the database every year. In 2010, this becomes two papers per minute. In fact, as [18] proved, the growth of MEDLINE is exponential.

Growth of knowledge and demand to access right and fresh information requires scientists to use tools so that they can more efficiently reach needed information and even uncover new relationships in that vast amount of data. That is where bioinformatics comes into play.

Bioinformatics is the field that combines computer science, information technology, and biology. Tools provided by bioinformatics help scientists analyze and interpret various types of data, including sequences of amino acids, numerical or textual data. Research areas in the field of bioinformatics include sequence analysis, genome annotation, literature mining, and analysis of many other biological subjects. Beside others, literature mining is the key area that deals with the analysis and interpretation of textual data and it is done by the help of the text mining methods.

3 Text Mining

Text mining is all about knowledge discovery. Unlike data mining, which requires highly structured data in mostly numerical format, text mining deals with the text that is explicitly expressed in the syntax and semantics of a natural language. Text mining involves detection, extraction, and maintenance of knowledge. To achieve this, it combines methods from computational and linguistic fields. In the next section, we will describe the elements of the Text Mining process. Then we will talk about the ways we use text mining in the context of bioinformatics.

3.1 Elements of Text Mining

As text mining mostly mimics the way database curator works, it starts with the determination of which resources to look at. These can be published papers, journal articles, patents, and even electronic medical records (EMRs). This step involves text classification and/or clustering, as well as information retrieval (IR). After selecting what to read, identification of important entities and relations between those entities in those selected documents are performed. These steps corresponds to name entity recognition (NER) and relation extraction in the context of text mining. After that, extracted information can be used for further analysis by summarizing the findings or answering questions specific to extract information. Following subsections describe each step.
3.1.1 Text Classification and Clustering

Basically, text classification and clustering is used to divide the collection of documents into smaller subsets based on their content. While classification is done by assigning pre-defined labels to each document, text clustering groups the documents without using any label in unsupervised fashion.

Since text classification and clustering techniques are very generic and can be used on any document, they have a long history in the literature. Methods used for text classification and clustering are borrowed from the machine learning field, such as naïve bayes classifiers, decision trees, maximum entropy models, and support vector machines.

For database curators, who are responsible for reviewing great numbers of documents to see whether or not they contain relevant information for their research, text classification tools have great importance. For example, during the Knowledge Discovery in Databases (KDD) Challenge Cup 2002, [25] tackled the problem of evaluating papers from FlyBase dataset and determining whether the paper should be curated based on the presence of certain experimental evidence. They achieved f-score of 78%. As in the case of this study, applying text classification methods for biomedical annotation and curation tasks is a new way of using such methods. For this purpose, in the Text REtrieval Conference (TREC) 2004, Genomic Track had a classification problem as one of its tasks.

3.1.2 Information Retrieval

One way of thinking text classification is to divide the documents into two classes: relevant and non-relevant based on given query. This process is called information retrieval. In fact, if you consider, all web search engines are information retrieval systems. They match given query with the documents in their collection and return most relevant ones in ranked order so that we can look though returned list of documents and obtain requested information fast. PubMed and Google Scholar are among the most known IR tools in the scientific community. To see more complete list of IR systems, check the Section 4.3.

How does an IR engine work? PubMed, which is the best-known biomedical IR system on the Internet, relies on two well-known IR methodologies; the boolean model and the vector model. The boolean model allows users to express their request in certain combination of terms and returns only those documents that contain those terms. If you enter a query like “yeast AND cell cycle”, it returns those documents that mention both yeast and cell cycle together. On the other hand, in the vector model, documents are represented as a vector in which each occurring word in the document is represented with a value calculated with a frequency-based weighing scheme. Like the documents, given query is also represented as a vector and what IR engine does is to compare the similarity between query vector and each document vector and return the results, ordered based on the degree of similarity. If you know what you are searching for, boolean model returns the exact results you are looking for. On the other hand, if you are not sure, then the vector model returns all similar matches to your loosely defined query. The choice depends you requirements.

Information retrieval systems are evaluated in terms of precision and recall metrics. Precision is the fraction of retrieved documents that are relevant, while recall is the fraction of relevant documents that are retrieved. To have a single metric, f-measure is also often used, which is harmonic mean of precision and recall. However, regarding the accuracy of these metrics, it is difficult to measure the efficiency of IR engines, especially since we cannot know the complete set of relevant documents to any possible query in the collection, calculating the recall of the system is not
plausible. Still, estimates done on most popular search engines such as Google suggests that even though it seems like we mostly satisfy with returned results, precision and recall values are measured below 0.3 on average, meaning that out of top 10 search results only 3 of them are relevant, while 70% of the relevant results are not even considered as relevant.

In the context of text mining tools, IR engines are mostly used as a stand alone component of the system. The output of the IR engine feeds the next component, which can be an QA or relation extraction engine. In case of the QA engine, returned results are further processed for a possible answer. In relation extraction case, it gets results at the sentence level and tries to match known patterns with returned sentences in order to extract any conveyed relation.

3.1.3 Name Entity Recognition (NER)

Name entity recognition is a task of identifying names of terms given in a text. Those terms include molecules, such as RNA, proteins, genes, diseases, chemical compounds etc. As scientific process is growing in great speed every day and more and more specialized subfields are added over the time, this results in lack naming conventions throughout the scientific community, which makes this task particularly difficult. In biology and medicine fields, the vocabulary of terms are large and very dynamic, exceeding many millions of terms. With large amount of information being generated by biomedical researchers, the terms that most suffer from synonym and homonym are gene and protein names. One perfect example is the \texttt{br} and \texttt{mod(mdg4)} genes of \textit{Drosophila} organism, which have 82 and 64 aliases, respectively.

Due to the rapid progress in genomic research in biomedical field, the identification of gene and protein names has evolved more than other types of terms in recent years. On the contrary, recognition of diseases and chemical compounds has attracted less attention, leading to poorer performance. Observations show that the more dictionary-based and manually crafted systems achieve best results, which is expected. However building such systems and keeping such dictionaries up-to-date requires tremendous amount of human effort, which is why today more and more machine learning based approaches are applied onto the NER task.

Methods used for NER fall into three categories: dictionary-based, rule-based and statistical approaches. Hybrid methods have also been used in the literature. Evaluation of applied methods are done with precision and recall metrics, as well as f-measure, just as in the case of evaluation of IR systems. The state-of-the-art system for human protein name recognition reaches around f-measure of 85% [32].

During recognition of name entities in the text, based on the requirements of given task, there can be various types of issues to be dealt with, such as text normalization, resolution of acronyms, as well as anaphoras.

**Term Normalization**

NER is to detect mentioned terms but one further step is to normalize them, which is to associate occurring terms with their corresponding entries in the database, or ontology, to be more exact. Ontology, which is also called knowledge base (KB), can be considered as a library of concepts and semantic relations between those concepts. Today one of the widely used ontology is Gene Ontology or GO in short, which describe the properties and functions of gene products in an abstract level.

The benefit of term normalization comes with the ontology because when we identify which concept of an ontology is mentioned in the text, we can also annotate that text with related
information defined in that ontology. For example, recognition of proteins from GO as well as protein residue and species allows us to annotate mutations with functional information.

**Acronym and Abbreviation Resolution**

One of the biggest sources of ambiguities in NER is the abundance of acronyms and abbreviations in the biomedical literature. One example is SCT, which is the official gene symbol for the human gene 'secretin.' However, most documents in MEDLINE using the item SCT may refer to one of the more than 100 different meanings, such as 'stem cell transplant', 'salmon calcitonin', or 'stair climbing test.' To human, this wouldn’t be any problem because he consider the terms within their surrounding context. However, in case of computers, since they cannot understand the context, it turns into an ambiguity. In order to do a better job in NER task, the systems have to deal with the resolution of acronyms and abbreviations. One way of doing this is to extract acronyms and then use that acronym vocabulary in your system, which is not a good way of tackling with this dynamic problem. Second way is to do automated term disambiguation.

**Anaphora Resolution**

Anaphora is an instance of an expression that refers to previously mentioned expression in the context. In a typical text, pronoun “it” is most used anaphora that has to be resolved, meaning that we have to identify, or disambiguate, which entity it refers to in the context and keep its record so that other tools which take that text as input can process that text without any ambiguity.

Anaphora resolution is considered to be a difficult task because getting it right requires to understand the meaning of not only sentence but also the context. Due to high complexity of NER task in the bioinformatics context, performance of the anaphora resolution system for this particular genre is expected to be lower than the performance of the same system for other genres.

**3.1.4Relation Extraction**

Identifying which terms are mentioned in the text is the first step towards understanding what it is saying. The next step is to extract relations between those terms occurring in a given sentence. In the literature, other than relation extraction, this task is also called fact extraction or, in more broader terms, Information Extraction (IE).

The simplest way of detecting relationships between entities is to look of co-occurring entity pairs in a sentence or a text. Because of two entities occur in the same sentence in number of times, it is a good indicator that those two entities are related to each other. Most of the early work on relation extraction was based on co-occurrence based approaches. Even though it is a very basic approach, if you have a large body of text, you can achieve high recall but with poor precision. Despite that, since they can easily identify any type of relationship, today co-occurrence based approaches are used as exploratory tools, especially by online web sites.

For the better part of the research community, co-occurrence based approaches now form the baseline of the relation extraction studies. They are replaced with pattern-based methods which provide much more precise and linguistic oriented solutions. These patterns can either be hard-coded patterns or automatically learned patterns from given examples, where machine learning comes into the play.

Recent studies are even searching for more sophisticated ways for extracting deeper semantics out of the sentences by making use of full parse trees generated by improved statistical sentence parsers, much bigger annotated biomedical corpora, and bigger ontologies for representing the
3.1.5 Text Summarization

Text summarization is a task of identifying the most important aspects of one or collection of documents and presenting these compactly and coherently. Based on the intended use of summary, automatic text summarization is done either generic or topic-specific, which is called targetted summarization in that case. Generic summaries can be indicative to help the reader make a decision about whether or not they need to read the whole document, or informative to present the told message in the document briefly.

Due to the complexity of the task, text summarization studies in the field of bioinformatics fields can be considered as slow compared to progress in the NER and IE tasks, but not as slow as the progress in the Question Answering task, which will be mentioned in the next section. In fact targetted summarization is quite similar to the question answering as it is done based on given query.

One of the studies on targetted summarization in bioinformatics field was done by [24] where they developed a method for generating structured summaries expressing six aspects of a gene: 1) gene production, 2) expression location, 3) sequence information, 4) wild-type function and phenotypic information, 5) mutant phenotype, and 6) genetic interaction. Summaries are generated from a set of MEDLINE abstracts retrieved via queries which are determined based on those listed aspects of the target gene. Evaluation of their system showed 50-70% precision in identifying these aspects of a gene.

In another study [5] in BioCreative 2006 sentence selection subtask of the protein-protein interaction task, they tried to find the best sentence that describes a protein-protein interaction in given text, which achieved a 19% correct rate. Even though their performance seems quite low, considering that the second-place system scored 6% somewhat indicates the difficulty of the task.

As summarization of a document can also be seen as compact representation of that document, another way of looking at this task is to provide better access to the facts, such as via better user interface design. New online tools such as iHOP, MedMiner, and Chilibot work on this aspect.

3.1.6 Question Answering

In [2], they looked at the average elapse time for a family physician to get an adequate answer to a clinical question and measured between 2.4 to 6.5 minutes. In another study [15], they observed that physicians spent less than 2 minutes on average to get an answer to a question. These studies clearly exemplifies that instead of searching through pile of documents via IR systems, we need a system that can deduce what we are looking for and give the exact answer to our questions. This is called question answering (QA).

Perhaps the first operational question answering system was MIT's START from 1993. Since then, techniques to process text in natural language has evolved considerably however the complexity of the question answering task remains to be unchallengeable, or should we say probably?

Maybe today's the most promising QA system is the Watson of IBM, which faced off against the two former champions of the TV quiz show called Jeopardy! on the national television in February of 2011. In the game, contestants are presented with clues in the form of answers in variety of topics such as history, literature, art, pop culture etc., and must phrase their responses in question form. Watson beat both champions by a wide margin, which was the second most anticipated
Despite Watson and handful of question answering engines, there hasn't been much research done on this intriguing subject in the field of biology and medicine. In his studies [35], Zweigenbaum analyzed the feasibility of question answering system in the biomedical domain. He talked about trustworthiness and availability of the knowledge on the web. He pointed out already available linguistic and terminological resources to built a open-domain question answering system. [26] followed up on that and introduced a open-domain QA system to answer genomic questions. MedQA is another QA system developed by [23] which makes use of information retrieval, extraction, and summarization techniques to answer definitional questions like “what is X?”. In yet another system, [34] made use of shallow parsing to look for sentences that are syntactically similar to given question and showed that it enhanced the performance of the answer extraction.

Despite these systems, the progress in the question answering task is very slow compared to other already talked tasks. Instead of building such very complex question answering systems, the trend in the research field is to enhance the IR system with better techniques so that it achieves higher precision and recall, which is eventually expected to reach the level of a question answering system. In that regard, instead of dealing with documents, retrieval is done at the passage or even at sentence level. These are also called advanced IR models. In one such study, [12] combined the results of four different IR systems and observed significantly better results. Others used language models and relevance feedback to boost the performance of their IR systems.

3.2 Usage of Text Mining in Bioinformatics

For scientists who are adrift in the digital ocean of knowledge, text mining is the genie that answers their most precious requests. Considering that every new day thousands of new reports and papers are added onto a gazillion of them, text mining help scientists dig up the information they need so that they can integrate that knowledge into their though process and even unearth new relationships along the way. Moreover, they can keep themselves up-to-date about new trends in their fields. Following subsections elaborates on these three major ways of using text mining tools.

3.2.1 Literature Mining

As its name suggests, literature mining is to detect and extract information out of text in the literature, such as papers and journal articles. In real world, this is the job of database curators who are the domain experts. They are responsible for determining what resources to collect, supervising their documentation, conducting research based on that collection, and sharing that knowledge with public and scientific community. As, nowadays, computers are more and more involved in such tasks, database curators are the most beneficiaries of text mining tools.

Most literature mining tasks target specific goal. In bioinformatics, these can be

- detecting interactions between
  - protein-protein
  - protein-gene
  - gene-disease
  - gene-drug
  - drug-drug
- finding sub-cellular localization of proteins
- annotation of proteins based on their functions
- discovering gene functions, relations, and functional similarities
- pathway discovery and
To capture each different type of information, various tools have been developed. While each one applies different approach, they all do text mining in essence.

3.2.2 Hypothesis Generation

Among the researchers and scientists, literature-based discovery is considered to be one of the potential sources of promising hypotheses. With every new paper, a new set of facts are introduced to the scientific community. If we can put some together in meaningful way and apply some logical inference, we may come up with a brand new hypothesis. This was what Swanson did.

In his studies, Swanson [29] realized the power of the transitivity relation between the facts, which is basically explained as “A influences B and B influences C, therefore A may influence C”. This is known as Swanson's ABC model or, in more technical terms, second-order relations. In 1986, he discovered a relation implying patient benefit between fish oil and Raynaud's syndrome [30]. This was two years before any clinical trial proved such relation. In another study, he discovered indirect associations between migraine and magnesium using only the summarization of published articles, which was again experimentally verified later on. Based on the success of Swanson, people tried to automate this hypothesis generation process. To give an example, [31] used co-occurring terms in the titles and abstracts of MEDLINE papers and discovered potential new uses of thalidomide. In another automated hypothesis generation study, [28] unearthed implicit relation between therapeutic effect of Curcuma longa on retinal diseases, Crohn's disease and spinal cord injuries.

All these studies indicate that extracting buried knowledge out of published papers and connecting them together by making use of logical inferencing exemplifies the most promising benefit of text mining because it is not only virtually impractical to read all papers as they publish but also virtually impossible to consider every possible way of connecting new facts with old ones and discover all possible hypotheses.

3.2.3 Trend Analysis and Mapping the Science

Another way of using text mining is to look for trends in the literature. Without requiring any deep understanding of the text, by just counting the number of times any term was mentioned each year, we can identify popular terms in the literature over the years. For example, we can look for most mentioned genes in last year's published papers.

Trend analysis not only summarizes what is hot or not in a period of time but also can give clues about the future. We can look for emerging buzzwords throughout the literature. For example, being able to see which proteins have a potential to be commercially hot next year can allow scientists to direct their study toward such promising topics and give them a chance to end up with scientifically and even commercially more productive research.

Along the same lines but in much bigger scale, we can even map the done science, meaning that we can identify so-far-tried methods in the literature, connect the articles and people who published them and predict the direction of the science for the future based on that charted map.

4 Resources

Since the first paper in the biomedical text mining was published in 1997, the community and the language resources have been evolved quite rapidly. Today, there are numerous text collections,
software modules and web-based applications for the biomedical text mining. In the following subsections, we will mention the most prominent ones.

4.1 Text Collections in Bioinformatics

The most prominent text collection for the text mining community in bioinformatics is, without a doubt, the MEDLINE which contains over 21 million paper abstracts. Even though abstracts are freely available, text mining community has no access to most of the full-text articles, especially of journal articles due to the copyright issues. Considering that full-text versions contain much more information compared to abstracts, it is the biggest setback for the community. Nevertheless, in some fields, such as chemistry, the situation is even worse, where even article abstracts are inaccessible.

<table>
<thead>
<tr>
<th>Full Text Corpora</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>HighWire Press</td>
<td><a href="http://highwire.stanford.edu">http://highwire.stanford.edu</a></td>
</tr>
<tr>
<td>PubMed Central</td>
<td><a href="http://pubmedcentral.org">http://pubmedcentral.org</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tagged Corpora</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>FetchProt</td>
<td><a href="http://fetchpod.sics.se">http://fetchpod.sics.se</a></td>
</tr>
<tr>
<td>GENIA</td>
<td><a href="http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA">http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA</a></td>
</tr>
<tr>
<td>PennBioIE</td>
<td><a href="http://bioie.ldc.upenn.edu">http://bioie.ldc.upenn.edu</a></td>
</tr>
<tr>
<td>Yapex</td>
<td><a href="http://www.sics.se/humle/projects/prothalt">http://www.sics.se/humle/projects/prothalt</a></td>
</tr>
</tbody>
</table>

Table 1: A non-exhaustive list of text collections

Table 1 above lists the most known text collections in the world of the biomedical text mining.

4.2 Terminologies and Ontologies

As already mentioned, the fields biology and medicine has a very large and dynamic terminology. Parallel to that, many resources have been developed to be able to contain all the terms in the literature. They are proven to be very useful tools for many tasks. To give couple of examples, in case of IR, they are used for query expansion. In case of basic NER, most known entities are kept in dictionaries so that it can detect them right away.

Terminologies can be in the form of taxonomies, controlled vocabularies, thesauri and dictionaries. For example, MeSH, is one of the widely used terminologies in the community. It is used to tag the scientific articles based on their content, which makes the IR easier. However, MeSH terminology is not exhaustive. To more accurately and completely represent any content, more detailed terminologies are required. This is where ontologies are needed.

With the development of Semantic Web and all semantic-based approaches, ontologies are the real crown jewels of this collection. They contain not only concepts but also definition of expected relations between those concepts. When occurring terms in the text are normalized and associated with their corresponding conceptual entities in ontologies, this allows us to annotate the text even with more detail and detect even more relations, especially helpful in case of hypotheses generation.
Table 2: A non-exhaustive list of ontologies, terminologies and their uses

<table>
<thead>
<tr>
<th>Ontology/Terminology</th>
<th>Used to Annotate</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO</td>
<td>Proteins</td>
<td><a href="http://www.geneontology.org">http://www.geneontology.org</a></td>
</tr>
<tr>
<td>IUBMB EC numbers</td>
<td>Enzymes</td>
<td><a href="http://www.chem.qmul.ac.uk/iubmb/enzyme">http://www.chem.qmul.ac.uk/iubmb/enzyme</a></td>
</tr>
<tr>
<td>Sequence Ontology</td>
<td>Sequences</td>
<td><a href="http://song.sourceforge.net">http://song.sourceforge.net</a></td>
</tr>
<tr>
<td>MIPS FunCat</td>
<td>Proteins</td>
<td><a href="http://mips.gsf.de/projects/functcat">http://mips.gsf.de/projects/functcat</a></td>
</tr>
</tbody>
</table>

Table 3: Biomedical Information Retrieval Systems

<table>
<thead>
<tr>
<th>Name</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ali Baba</td>
<td><a href="http://alibaba.informatik.hu-berlin.de/">http://alibaba.informatik.hu-berlin.de/</a></td>
</tr>
<tr>
<td>BioContrasts</td>
<td><a href="http://biocontrasts.biopathway.org">http://biocontrasts.biopathway.org</a></td>
</tr>
<tr>
<td>BioText</td>
<td><a href="http://biosearch.berkeley.edu">http://biosearch.berkeley.edu</a></td>
</tr>
<tr>
<td>Citexplore</td>
<td><a href="http://www.ebi.ac.uk/citexplore/">http://www.ebi.ac.uk/citexplore/</a></td>
</tr>
<tr>
<td>EBI Med</td>
<td><a href="http://www.ebi.ac.uk/Rebholz-srv/ebimed/">http://www.ebi.ac.uk/Rebholz-srv/ebimed/</a></td>
</tr>
<tr>
<td>eTBLAST</td>
<td><a href="http://invention.swmed.edu">http://invention.swmed.edu</a></td>
</tr>
<tr>
<td>Google Scholar</td>
<td><a href="http://scholar.google.com">http://scholar.google.com</a></td>
</tr>
<tr>
<td>GoPubMed</td>
<td><a href="http://www.gopubmed.org">http://www.gopubmed.org</a></td>
</tr>
<tr>
<td>HubMed</td>
<td><a href="http://www.hubmed.org">http://www.hubmed.org</a></td>
</tr>
<tr>
<td>iHOP</td>
<td><a href="http://www.ihop-net.org/UniPub/iHOP/">http://www.ihop-net.org/UniPub/iHOP/</a></td>
</tr>
<tr>
<td>Info-PubMed</td>
<td><a href="http://www-tsujii.is.s.u-tokyo.ac.jp/info-pubmed/">http://www-tsujii.is.s.u-tokyo.ac.jp/info-pubmed/</a></td>
</tr>
<tr>
<td>McSyBi</td>
<td><a href="http://textlens.hgc.jp/McSyBi/">http://textlens.hgc.jp/McSyBi/</a></td>
</tr>
<tr>
<td>MedEvi</td>
<td><a href="http://www.ebi.ac.uk/tc-test/textmining/medevi">http://www.ebi.ac.uk/tc-test/textmining/medevi</a></td>
</tr>
<tr>
<td>MedMiner</td>
<td><a href="http://discover.nci.nih.gov/textmining">http://discover.nci.nih.gov/textmining</a></td>
</tr>
<tr>
<td>MEDIE</td>
<td><a href="http://www-tsujii.is.s.u-tokyo.ac.jp/medie/">http://www-tsujii.is.s.u-tokyo.ac.jp/medie/</a></td>
</tr>
<tr>
<td>PubMed Central</td>
<td><a href="http://www.pubmedcentral.nih.gov/">http://www.pubmedcentral.nih.gov/</a></td>
</tr>
<tr>
<td>Textpresso</td>
<td><a href="http://www.textpresso.org/">http://www.textpresso.org/</a></td>
</tr>
<tr>
<td>Twease</td>
<td><a href="http://twease.org/medline/app">http://twease.org/medline/app</a></td>
</tr>
</tbody>
</table>

Table 2 above lists some of the terminologies and ontologies used in the biomedical text mining community. It is most certain that over the upcoming years more and more such resources covering much more concepts will emerge.

4.3 Online Tools

The final product of the text mining is extracted information out of free text in natural language and IR and QA systems are the tools that present that information in various formats. Even though QA systems are not there yet, there are full-fledged online IR systems on the net. Table 3 lists the most known of those IR systems.
5 Conclusion: The Future

For text mining in bioinformatics, the future is coming fast. Even though it is a very new field, the motivation behind this progress is very strong, fueled with the importance of health care in our lives, as well as prospective commercial aspects of the field. Following discussion elaborates on each issue that will come up in the future of the text mining in bioinformatics.

5.1 Wanted: More Resources

As more and more applications rely on machine learning methodologies, in order to train such systems and getting the best performance out of them, we are going to need much more labeled data. However this is not only the problem of bioinformatics. Labeled data is the most valuable piece of resources in the literature. To remedy this, more studies are done on unsupervised learning techniques, which requires less or no labeled data.

Still, we need labeled data not only for learning but also for evaluations. Today's widely used labeled corpora came out of the evaluation competitions like KDD Challenge Cup, BioCreative I and II. Better evaluation means better understanding of the positive and negative sides of the methods, which eventually leads to development of better methods.

In addition to the need for labeled resources, there is much more pressing concern in the biomedical text mining community, and that is not accessible full-text articles. So far, more than 90% of the biomedical text mining has been done on the MEDLINE abstracts. Considering the rapid growth of the text mining community with just abstracts, we can stop asking to ourselves what would happen if they have the whole text. Not only there will be more chance to tune up their methods, but also online information retrieval tools will be able to return more relevant articles with higher precision and recall.

5.2 Wanted: More Research on Text Summarization and Question Answering

Even though advanced IR systems steal some attention from QA systems, the future are going to need only one answer. IR systems that perform at high quality can only go to certain level and they will still present us a list of documents at the end. However, QA systems is the real and the only answer for the future.

Today's QA engines might be seen naïve as they are combination of IR + IE + TS. To achieve higher level, we need to study more on the understanding of the natural language by working at the semantics level. We need to leave the bag of words and co-occurrences models and try to represent the information in terms of well defined concepts and relations. Initiatives that support the construction of ontologies and knowledge-bases are the first steps we can take on this road. Gene Ontology (GO) is the perfect example.

Moreover, we need our question answering systems act like expert systems so that we can get into the conversation with them. Learning something one question at a time might not be enough. We need systems that follow up on the discussion so that it can understand the needs of the user better.

Even though it is expected that QA systems are taking their first steps today, slow pace in text summarization field in bioinformatics is quite surprising. Even though it is related to the complexity of the field of bioinformatics, in newswire genre, there have been quite number of studies done on text summarization. Before achieving full-fledged QA systems, we first have to tackle with the text
summarization task, especially targeted summarization because identifying what is important in the context is something that QA system needs for better understanding of the content.

5.3 Wanted : Integration Frameworks

One of the important issues for the text mining efforts is that even though extracted knowledge is quite valuable, being able to connect such knowledge with biological and biomedical data makes that knowledge even more valuable. We need tools that not only answer the questions of physicians and practitioners but also access to those biomedical data and put those into use in the light of knowledge they have. Not only that, being able to combine text and biological data leads to better hypothesis generation and new discoveries.

5.4 Wanted : User-centric Tools

In order to create better tools, first we need to understand the needs of the users. So far text mining community spend more time with the text than the real users who are going to use their final product. This is exemplified with one study which showed that among all online tools that are available on the Internet, the most used ones are those that are developed by real scientists.

Studies on the needs, behaviors and interactions of users with the tools are an effective way of identifying the requirements of those desired tools. One such study [22] investigated the information-seeking behavior patterns. Their study was based on previous study [14] done on researchers and their way of using information retrieval systems and found out that information-seeking behavior can be characterized with only small number of distinguishing types of activities: starting, chaining, browsing, differentiating, monitoring, and extracting. When [22] applied this approach onto the biomedical context, they realized that there are tools for every activity but not much for differentiating the data. For example, online tool BioContrasts [21] looks for the contrastive relations between proteins by using language patterns like “A but not B”

In the future, we need more systematic and user-centric evaluations of the tools so that we can better understand the needs of the real users and develop better tools at the end.

5.5 Wanted : Community Involvement thru Social Networks

If today we are living in the Internet era, right now is the era of the social networks. Thanks to the connective power of the Internet, thousands of millions of people are interacting with each other every day. People share their thoughts, pictures, even their whole social lives. What comes out of this is the productive power of communication. Even though most of the social transactions are considered to be garbage, using the same tools to let scientific community interact for good is the challenge that we have to embark.

One perfect example for such case is the wikis. A wiki is a website that allows people to create and edit content collaboratively. The most known example is the Wikipedia, which is an online free encyclopedia created collaboratively by hundreds of thousands of people. Allowing people to publish their findings thru wikis or let them annotate resources collaboratively is a great way of generating more resources for the text mining tools.
References


Cell; 21: 589-94


