Cell

Seeking a New Biology through Text Mining

Andrey Rzhetsky,[1,*] Michael Seringhaus,[2] and Mark Gerstein[2]
[1]University of Chicago, Chicago, IL 60637, USA
[2]Yale University, New Haven, CT 06510, USA
*Correspondence: arzhetsky@uchicago.edu
DOI 10.1016/j.cell.2008.06.029

Tens of thousands of biomedical journals exist, and the deluge of new articles in the biomedical sciences is leading to information overload. Hence, there is much interest in text mining, the use of computational tools to enhance the human ability to parse and understand complex text.

Imagine that a graduate student enters the U.S. Library of Congress with the goal of retrieving all texts relevant to protein glycosylation. Her problem is straightforward, known among text miners as *information retrieval* (IR). If the student must not only find the books but also flag the most important concepts she encounters in each, she is performing *named entity recognition* (NER). Undaunted by her workload, imagine she decides to identify relations between concepts, such as "protein BAD binds to protein BAX" (called *information extraction* or IE). Then she takes on additional tasks such as question/answer (QA) and text summarization (TS). Computational IR, NER, IE, QA, and TS are all part of text mining and belong to the larger field of *natural language processing* (NLP), which itself is a part of *artificial intelligence* (AI) that aims to recreate or surpass the computational ability of the human brain. Although multiple definitions exist, text mining is typically associated with information retrieval, extraction, and synthesis, with a special emphasis on gaining new knowledge (Table 1).

The Elements of Text Mining

Bringing the Modules Together

Text mining has an established formula for joining the computational and linguistics modules together (IR + NER + IE + QA + TS). The first step is identification and retrieval of relevant documents (IR) (although this step is optional and is sometimes substituted with indiscriminate analysis of documents regardless of their relevance). Once documents are available for computational analysis, an NER module is put to work, followed by an IE module to extract relations between entities. The extracted nuggets of information can then be used for TS, QA, and a higher-order analysis capable of proposing new conclusions.

To retrieve information properly, an application has to "know" the relations and entities mentioned in the documents it is searching. Therefore, the boundary between NER, IR, and IE is fuzzy and can be encapsulated in an application that attempts to do all three tasks jointly. Here, we present some of the key considerations common to any text-mining approach, discuss how these relate to the changing landscape of scientific information, and give an overview of current and future applications of text mining to the scientific literature.

Information Retrieval (IR). Every scientist is familiar with IR: we use various incarnations of IR when we conduct computer-aided searches for articles, books, and Internet sites. Indeed, PubMed and Library of Congress searches and Google Scholar are among the most visible IR instantiations. The ideal IR system should be able to deduce what we are looking for, even if we are somewhat fuzzy in the question we pose. It should retrieve and rank the results by their relevance to our question. It should also extract text-encoded facts and compactly summarize them, as a capable and tireless assistant would do.

It is difficult to benchmark the efficiency of IR engines, especially their recall, because the complete set of documents relevant to almost any search is inherently ill defined. Nevertheless, estimates show that the most popular search engines, such as Google, have both precision and recall below 0.3 (Shafi and Rather, 2005). In other words, every time we do a search, more than 70% of the documents in the output are irrelevant, whereas more than 70% of all relevant documents never appear in the engine's output.

Information Extraction (IE). To "comprehend" text, a computer program has to map words and phrases to objects, concepts, and symbols. For example, if the program encounters the phrase "pray for elves," it must decide if these words refer to separate entities (e.g., a gene named "elves"), an English directive, a single gene with a three-word name (the eponymous gene does exist in the fruit fly *Drosophila melanogaster*), or something entirely different. Similarly, when it encounters the sequence of characters "Alzheimer," it needs to "decide" if the word refers to Dr. Alois Alzheimer who first described Alzheimer's disease or

Table 1. Text Mining Web Resources

BLIMP (Biomedical Literature-Mining Publications)	http://blimp.cs.queensu.ca/
Alexander Morgan's compilation of BioNLP resources and references	http://compbio.uchsc.edu/corpora/bcresources.html
Resource links compiled by Dietrich Rebholz-Schuhmann	http://www.ebi.ac.uk/Rebholz/resources.html
Text-mining resources compiled by Robert Futrelle	http://www.ccs.neu.edu/home/futrelle/bionlp/
A list of links to current NER, IR, and IE engines	http://www.bork.embl.de/Docu/literature_mining/
Marti Hearst's What Is Text Mining?	http://people.ischool.berkeley.edu/~hearst/text-mining.html

to the disease itself. This problem in text analysis is typically referred to as named entity recognition (NER) and has proved particularly challenging for biomedical prose. Although NER engines that mine news articles regularly achieve accuracy above 90%, their biomedical counterparts are more likely to perform at 80% or below (text miners typically compute the F-measure rather than accuracy) (Cohen and Hersh, 2005; see Table S1 available online). This is because biology and medicine are unusually rich in terminology: the collective vocabulary used by biomedicine incorporates many millions of terms. The exact number is unknown and constantly in flux. Because this vocabulary is large and dynamic, new terms emerge rapidly and erratically. As a result, the same real-world object may have numerous names (synonyms), whereas distinct objects can be identified with the same name (homonyms). The terms that most notoriously suffer from synonym and homonym abundance are gene and protein names (Hirschman et al., 2002; Wilbur et al., 1999). A given gene may be denoted by several dozen synonymous names—for example, the *Drosophila* genes *br* and *mod*(*mdg4*) have 82 and 64 aliases, respectively. Even worse, vastly different and incompatible naming systems are used in different species, and gene aliases themselves merge into intricate semantic networks that connect gene names, pop-culture catch phrases, idioms, and common everyday utterances borrowed from multiple languages.

From Co-occurrence to Deep Parsing. Imagine that you have to find out as much as possible about processes involving glycosylation—the enzymatic attachment of sugars to proteins and lipids. Spotting and listing most mentions of glycosylation in the literature is an unbearably tedious task for a human, not because any single mention is hard to find or understand, but because there are millions of biomedical articles, many of which describe glycosylation. Imagine now that you seek only four pieces of information for each glycosylation description: which molecule is being glycosylated, at which position, by which enzyme, and what kind of sugar is being added. Thus, the automated extraction of the following quartet is a perfect job for a typical IE engine.

Glycosylate (What, at Which Position, by Which Enzyme, with Which Sugar)? Unlike a human curator, an IE engine would not capture the nuances of individual articles, such as an out-of-date experimental procedure or a questionable connection between experimental results and article conclusions. However, an IE engine could process enormous volumes of texts without succumbing to fatigue or ennui. To perform tasks like this, simple IE engines use only the co-occurrence of terms in the same sentence, paragraph, or article (with a statistical filter superimposed to distinguish coincidence from significant hits). Such co-occurrence IE engines are typically fast but not very precise. Increasing sophistication (and precision) in IE engines usually involves borrowing methods from computational linguistics. For example, so-called chunking (shallow parsing) can help us to identify syntactically related groups of words (such as noun and verb phrases) and subsequently improve downstream processes, such as NER.

The more challenging slice of the IE engine spectrum (deep parsing) is built on formal mathematical models describing how text is generated in the human mind, the so-called formal grammars. The most popular formal grammars are deterministic or probabilistic context-free grammars (Ananiadou et al., 2006; Table S1). Grammar-based IE is computationally expensive because it requires the evaluation and ranking of a large number of alternative ways to generate the same sentence (alternative parses); it is therefore considerably slower but potentially much more precise.

Synthesizing and Finding New Connections. Moving beyond information retrieval and extraction, some proportion of published assertions can be repackaged to form "synthetic ideas," that is, new compound concepts that are significantly more valuable to the scientific community than the sum of their original assertions.

Even with current text-mining capabilities, such synthetic ideas can be discovered automatically. A more distant but nonetheless realistic aim of the field is to trace and map more sophisticated ideas (idea isomorphisms) that are expressed differently in different scientific fields yet represent identical problems or their solutions. If such idea mappings were made instantly available through an Internet interface, the result could be truly impressive. The diffusion of innovations across science could be markedly increased by making solutions developed in one area visible to specialists still searching for them in a different field. Computationally pairing problems and solutions generated by different fields is a type of automated creativity (systematic search for synthetic ideas) that computers almost certainly will do for us in the not too distant future.

Question Answering (QA). Information retrieval/extraction and summarization are often married in a single application, a question-answer system. Perhaps the first operational question-answer system was START, designed and launched at the Massachusetts Institute of Technology in 1993 (http://start.csail.mit.edu/). START remains operational today and serves as a general encyclopedia (try it!). But there are limitations with current question-answer systems. You can ask START to list all cities within a 250 mile radius of Chicago (or another city), which it does promptly, listing the actual distance between Chicago and every city in the list. START, however, has only limited knowledge about molecular biology. With current tools, given a large collection of molecular interactions, it would be relatively easy to write a similar engine that answers biological questions like "list all molecular interactors of protein X with property Y." Such a question-answer system would surpass abilities of both human experts and currently available encyclopedias.

Text Mining and the Structure of Scientific Information

Evaluating Tools: Benchmarking

Technological advances demand quality control. In text mining, we can ask: how effective are the tools that are currently available? Text miners typically evaluate their own methods, but it is difficult to evaluate rival methods applied to different test data (Krallinger and Valencia, 2005; Jensen et al., 2006). These considerations have spurred competitions where researchers can pit differ-

ent methods against one another while processing identical data sets. The KDD Cup (Knowledge Discovery and Data mining), TREC Genomics (Text Retrieval Conference – Genomics track), and BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) are examples of such contests (see Table S2). These competitions share a similar underlying design. Organizers choose a well-defined technical task, such as the identification of a gene or protein name embedded within a collection of text snippets. The competitors—research groups around the globe—receive the same challenge, test data, and information. To evaluate the participants' success, the organizers recruit a group of experts that judge the participants' individual submissions or perform the competition task themselves to manually generate a gold standard against which to compare the computational submissions.

Aside from their obvious value in advancing text-mining techniques, these competitions have an important ancillary outcome: they offer the community a glimpse of the uneven complexity of various test-annotation tasks for human experts. Even well-defined tasks like normalizing gene and protein names are easier in some model organisms (e.g., yeast) than others (e.g., fly, mouse, and human) (Colosimo et al., 2005). Agreement among expert annotators varies from 90% to 70% depending upon the species considered. Thus, the perceived performance of automated tools strongly depends upon the test data selected, and when really tough tasks are considered, even human experts sometimes cannot agree.

The Changing Landscape of Journals and Databases

Text mining has important implications for the structure of scientific publishing itself. In the past decade, electronic databases have revolutionized the way that scientists access information. Today, digital repositories accommodate vast amounts of data, storing primary and ancillary records, functional annotation, and conservation information. Equally important, they allow updates to stored information, a luxury routinely exercised as sequences are corrected and revised.

The current format of scientific journals follows a model established long before the era of computers, cheap electronic storage space, and digital publishing. This arguably outdated format limits scientific communication. Ideally, scientists should record and share all useful findings, but in reality results sometimes do not coincide with the "standard ration" suitable for journal publication. Some facts are simply too trivial to merit a paper, and isolated findings or negative results are often withheld from the published record. Conversely, some data sets are too large to include in the text-based article format; for example, in two studies of whole-organism protein-protein interaction networks or regulatory pathways, the manuscripts presented highlights of the results and discussion whereas the data sets themselves are stored in databases or on laboratory websites.

By contrast, scientific databases are highly structured and machine readable. They require significant effort to establish and resist changes in focus because of the rigidity of their structure. Moreover, databases often lack true peer review, and because no uniform citation system exists to track the database contributions of a given researcher, there are few incentives to populate, annotate, or revise information stored in databases. Also, databases have not yet been optimized for discussion, to allow disagreement, or to represent uncertainty.

Thus, journals and databases are historically positioned to handle different types and amounts of data. Journal articles are optimized for human consumption and incorporate authoritative peer review. However, they are not suited to handle very large or very small results, they lack the consistent and rigid data structure required for easy computational access, and third-party indexing of full-text material is difficult.

Structured Texts and the Semantic Web

The roles of journals and databases are blurring (Bourne, 2005) as articles are accessed increasingly through database-type portals, and databases store article-like textual data (Hamosh et al., 2000). Although it is still impossible to compute with unstructured text as easily as with structured databases, there is a possible combination of community efforts and computational advances that can help to bridge the gap called the semantic web (Berners-Lee and Hendler, 2001). The semantic web is a new iteration of the World Wide Web in which a formal representation of the *semantics* (meaning) of each piece of information is provided along with the information itself, allowing computers to reason across text in human-like fashion. The semantic web can be particularly useful in connection with text mining, as it provides a way to mark up text with systematic and structured meta-information.

Text miners hope to convince publishers to enrich the plain text of manuscripts with computer-readable annotations. For example, the authors could be required to annotate gene, protein, and disease names within the text with a set of journal-specified tags (in the same way as a new gene sequence needs to be submitted to a central database prior to manuscript publication). The text-mining community could provide publishers with a set of tools that would automatically pre-annotate manuscripts before publication, subject to approval of authors (Seringhaus and Gerstein, 2007). This change undoubtedly may be hard to implement but would benefit science because the resulting semantic-web-enabled articles would be much easier to use both for information retrieval and extraction, as well as for the generation of knowledge. Researchers will be able to extract more value from a data set by analyzing experimental data jointly with textual information, for example, the analysis of full-text articles in combination with gene expression data (Natarajan et al., 2006) and the prediction of the subcellular localization of new proteins (Shatkay et al., 2007).

Applying Text Mining to the Scientific Literature

Application 1: Knowledge Is Zipfean

Frequencies of word use in everyday English (and other languages) follow Zipf's distribution, also known to economists as the Pareto distribution of wealth among people and to physicists as the power law. In a nutshell, Zipf's law in linguistics states that there is a very small subset of words that occur very frequently ("rich words") and a

large subset of words that are rare ("poor words").The same Zipfean regularity may operate over frequencies of statements and arguments that people use in everyday reasoning and in scientific thinking. There is a tiny set of facts that is known to virtually everyone (think about the "rich" memes *double helix* and $E = mc^2$), and there is a rapidly growing number of facts that are known to smaller and smaller groups of people.

Application 2: Consistency of Data

An experienced biologist can readily spot inconsistencies in a small map of molecular interactions. The biologist looks for logical inconsistency across multiple statements that are each perfectly reasonable when viewed in isolation but that cannot all be true simultaneously. For example, consider the problem inherent in accepting as true the following three statements about genes *A* and *B*: "*A* inhibits *B*," "*B* inhibits *A*," and "*A* and *B* are both active simultaneously." Such simple problems are easy to spot, but as a pathway model grows, tracking down inconsistencies becomes progressively more difficult. For that task, we will need an appropriately trained pathway tool similar to the common spellchecker that tags conflicting pieces of information with a different degree of certainty and then finds the most likely resolution of inconsistencies. This approach would not replace data analysis by biologists but may save time in refining models (just as a good spellchecker does not replace a writer but allows us to produce prose free of typos faster).

Application 3: Charting the Development of Science

Now that essentially all of science is recorded in electronic journals and online data, a new type of historical analysis is possible. One can use text mining not only to look for new connections in different fields but also to study the structure of science itself. Most dramatically, one can watch the birth of a new field by observing time slices of databases. For instance, it is possible to see the birth of the field of RNA interference from its 1997 conception to the present. We can also study the overrepresentation of certain fields relative to others, such as the preponderance of Nobel prizes in a particular field (e.g., crystallography) compared with the number of papers on the topic stored in public databases. Such approaches enable us to study the ways in which scientists develop and transmit ideas. We can also chart maps looking at the broad interconnections between different scientific fields (Shiffrin and Borner, 2004).

Through text mining, one can also study the different structures of scientific collaborations, contrasting those that are close-knit with others that are more disparate. Eventually, such meta-analysis of publication structure could help to determine how to structure large-scale collaborations. To some degree, this is already underway with citation analyses that look for publications and authors that are cited many times over.

Application 4: Is the Latest Word the Most Accurate?

Many handcrafted (as opposed to automatically generated) knowledge bases operate under the assumption that the most recently published facts are the most reliable. Although this assumption is reasonable in many (even most) cases, there are exceptions including errors made by leaders in the field and occasionally the publication of fabricated facts. With automated text-mining tools, we can keep in our analysis all instances of contradictory statements related to the same issue, ordered chronologically and linked to their textual sources.

Statements about the same problem published at different time points are by no means independent: researchers use prior publications in interpreting their experiments, thus introducing unwilling bias. Once the published statements are unlocked from their textual shells and deposited into computer-accessible media, they can be used for statistical analysis. Using a large collection of text-derived facts and a proper statistical model, we could potentially use computation to deconvolute errors associated with the subconscious bias that results in gradual distortion of experimental results in publications. It should be possible, both in terms of modeling and computation, to assign a quality value (such as the probability that the statement is true) to every published statement in the scientific "bibliome."

Conclusion

The process by which we acquire instruments for our intellectual toolbox—the facts and beliefs that we use in making inferences about the world—resemble in a sense the maturation of our immune system, which early on learns to distinguish self from non-self. A similar process of intellectual maturation allows us to streamline our thinking and avoid questioning every step of our mental process anew; the downside, however, is that, like the immune system, our early implanted antigens (or, beliefs) can later prove harmful by forming blind spots in our immune or mental landscape. Text-mining tools, which merge the precision and data handling of the computer with nascent technologies to parse human language, can synthesize information across broad fields of inquiry and may one day provide a way to lift the veil from the blind spots in our thinking.

Supplemental Data

Supplemental Data include two tables and can be found with this article online at http://www.cell.com/cgi/content/full/134/1/9/DC1/.

ACKNOWLEDGMENTS

We thank M. Çokol, A. Divoli, L. Dupré Oppenheim, T. Conrad Gilliam, I. Iossifov, R. Friedman, R. Rzhetsky, K. Smith, and C. Weinreb for helpful comments and L. Hirschman, N. Smallheiser, H. Shatkay, and W.J. Wilbur, for advice and providing references to publications and websites. This work was supported by the NIH (GM61372) and Cure Autism Now Foundation (to A.R.) and the Keck foundation (to M.G.).

REFERENCES

Ananiadou, S., Kell, D.B., and Tsujii, J. (2006). Trends Biotechnol. *24*, 571–579.

Berners-Lee, T., and Hendler, J. (2001). Nature *410*, 1023–1024.

Bourne, P. (2005). PLoS Comput. Biol. *1*, 179–181.

Cohen, A.M., and Hersh, W.R. (2005). Brief. Bioinform. *6*, 57–71.

Colosimo, M.E., Morgan, A.A., Yeh, A.S., Colombe, J.B., and Hirschman, L. (2005). BMC Bioinformatics *6* (*Suppl 1*), S12.

Hamosh, A., Scott, A.F., Amberger, J., Valle, D., and McKusick, V.A. (2000). Hum. Mutat. *15*, 57–61.

Hirschman, L., Morgan, A.A., and Yeh, A.S. (2002). J. Biomed. Inform. *35*, 247–259.

Jensen, L.J., Saric, J., and Bork, P. (2006). Nat. Rev. Genet. *7*, 119–129.

Krallinger, M., and Valencia, A. (2005). Genome Biol. *6*, 224.

Natarajan, J., Berrar, D., Dubitzky, W., Hack, C., Zhang, Y., DeSesa, C., Van Brocklyn, J.R., and Bremer, E.G. (2006). BMC Bioinformatics *7*, 373.

Seringhaus, M.R., and Gerstein, M.B. (2007). BMC Bioinformatics *8*, 17.

Shafi, S.M., and Rather, R.A. (2005). Webology 2, Article 12. http://www.webology.ir/2005/v2n2/a12.html.

Shatkay, H., Hoglund, A., Brady, S., Blum, T., Donnes, P., and Kohlbacher, O. (2007). Bioinformatics *23*, 1410–1417.

Shiffrin, R.M., and Borner, K. (2004). Proc. Natl. Acad. Sci. USA *101* (*Suppl 1*), 5183–5185.

Wilbur, W.J., Hazard, G.F., Jr., Divita, G., Mork, J.G., Aronson, A.R., and Browne, A.C. (1999). Proceedings / AMIA Annual Symposium, 176–180.