Contents lists available at ScienceDirect

Applied & Translational Genomics

journal homepage: www.elsevier.com/locate/atg

Will the future of knowledge work automation transform personalized medicine?

Gauri Naik [a,*], Sanika S. Bhide [b]

[a] Optra Health, 530 Lakeside Drive, Suite 250, Sunnyvale, CA 94085, United States
[b] Optra Systems, 503, B-Wing, Manikchand Icon, Dhole Patil Road, Pune 411001 India

ARTICLE INFO

Article history:
Received 11 April 2014
Received in revised form 16 May 2014
Accepted 19 May 2014

Keywords:
Genomics
Computational intelligence
Machine learning
Natural language processing
Automation

ABSTRACT

Today, we live in a world of 'information overload' which demands high level of knowledge-based work. However, advances in computer hardware and software have opened possibilities to automate 'routine cognitive tasks' for knowledge processing. Engineering intelligent software systems that can process large data sets using unstructured commands and subtle judgments and have the ability to learn 'on the fly' are a significant step towards automation of knowledge work. The applications of this technology for high throughput genomic analysis, database updating, reporting clinically significant variants, and diagnostic imaging purposes are explored using case studies.

1. Background

2008 was a landmark year in the history of human genomics research, when the pace of advances in genome sequencing technology exceeded Moore's Law (Mardis, 2011). This was the beginning of 'disruptive technology' in the context of furthering our understanding about the basis of being human. Since then, genomic science has never looked back, and every day scientists around the world make new dents in furthering this discipline. However, these advances are not exempt from impeding and ever evolving questions, which may call for a paradigm shift.

Several notable advancements in human genomics were made during the last decade, primarily in the area of improving DNA sequencing methods to significantly reduce run time (Tucker et al., 2009), increase accuracy of reads (Quail et al., 2012) and decrease cost towards sequencing 1 million base pairs (Liu et al., 2012). In the current and coming decade however, the major challenges would involve developing technologies that utilize the glut of data output from these sequencing platforms towards actionable insights.

On a high level, these future technologies will have to address core knowledge work activities encompassing everything from retrieving, storing, organizing, evaluating, and analyzing data to presenting aggregated information in the form of usable facts. This will underscore the beginning of an era when automation of knowledge work will change the pace of genomic research.

2. Automation of knowledge work

Why think about automation of knowledge work? Because, in today's rapidly moving world, many processes in diverse areas are already getting automated. To name a few, ATMs, credit cards, air traffic control, self-serve airline reservation systems, tasks such as taking deposits and checking customers out of a grocery store and many other manufacturing and transaction work are now automated (Manyika et al., 2013). Currently available intelligent personal assistants such as Apple's Siri 9 http://www.apple.com/ios/siri/ and Google Now http://www.google.com/landing/now/ are best examples of automated technologies in administrative support roles. Siri allows user to use voice to send messages, schedule meetings, place phone calls etc. It understands what we say, knows what we mean and even talks back. Google Now works in background and gives the right information at the right time. It gives weather information, plans the best route to avoid traffic, checks the favorite team's score when they are playing and much more. However, these applications work mainly as decision support tools on the basis of functionalities like search, analysis, sense-making and recommendation. Thus, these are examples of automating tasks that are routine or laborious and of value. Automation of routine knowledge work also offers a premise of anywhere access to intelligent and expert tools to support non-experts. There is still a significant gap to unite technology, information and human ability to understand and utilize the generated data, in a unified platform. We are yet to begin automating tasks that are custom, need significant common sense or contextual knowledge and are built on human relationship skills.

[*] Corresponding author. Tel.: +91 408 899 7390, +91 408 621 2195 (mobile); fax: +91 978 268 5227.
E-mail addresses: gauri@optrasystems.com (G. Naik), s.bhide@optrasystems.com (S.S. Bhide).
URL: http://www.optrahealth.com (G. Naik).

http://dx.doi.org/10.1016/j.atg.2014.05.003

Healthcare and biomedical industry are the most important domains where rapid advances in second generation automation of knowledge work may have significant impact. At the same time, as there are very few routine tasks in biomedical or healthcare industry compared to others, it will also be a challenging feat to achieve. Specifically in the context of genomic research, automation of complex workflow tasks, coupled with the ability to make specified autonomous decisions, can advance this scientific field by conferring unprecedented productivity where it is most required. With the technologies like high throughput screening (HTS) and next generation sequencing (NGS), a massive amount of biological data is getting generated. HTS can quickly conduct a large number of pharmacological, chemical or genetic tests, Next generation sequencers can sequence RNA transcriptomes to deliver extraordinary, high definition views of transcript sequence, SNP haplotypes, rare variants, splicing, exon boundaries and RNA editing (Toma et al., 2011). To keep pace with massive speed and data overload of these systems, we need the technologies to swiftly sift through already known facts, and use this information to 'identify' as well as 'tag' genes, antibodies or active compounds of importance in real time data. To accomplish such varied and complex tasks, knowledge work automation will have to steer clear of the pattern followed by evolution of factory automation and instead pave way for intelligent software and algorithms that have the ability to 'learn on the fly'.

2.1. Intelligent software and algorithms: the pillars of automation of knowledge work

Artificial intelligence (AI) algorithms have benefited fields such as law and financial services by knowledge work automation. These algorithms are able to parse multiple news stories, financial announcements, and press releases, make decisions regarding their trading relevance, and then act on it faster than any other human trader (Manyika et al., 2013). Similarly machine learning is best in complex analytics. Machine learning techniques such as deep learning and neural networks are fundamental implementers of knowledge work automation. Large amounts of time are spent in literature search in life science domain. To reduce the time required to extract the needed information from the articles, several methods for automated knowledge extraction are being developed; to cull information from biomedical literature, based on syntax trees and natural language processing. One such example is SENNA 'Semantic Extraction using a Neural Network Architecture' which uses semantic labeling of relation using verb-candidates (Barnickel et al., 2009). Similarly, robust engines for analyzing and managing unstructured data by using statistical methods such as latent semantic indexing (LSI), Bayesian modeling and neural network approaches are also being developed to augment automation of knowledge work (Chen et al., 2013).

2.2. Automation of knowledge work: a case study of genomics literature annotation

The literature in the field of genomics is increasing greatly with each passing day. The volume of new literature each year, measured in bytes, is about fifty times the size of the entire human genome. But locked in this literature is an enormous amount of information that can tell us much about the structure and function of genes, proteins, cells and organisms how they work as well as how they can fail. To make full use of information contained in such studies for diagnostic and prognostic purposes, it is imperative that significant findings be annotated to generate a rich database of biological entities and relationships.

Automation for capturing information from genomics studies is pertinent given the rapid increase in the rate of publication, as reflected in the growth in the contents of PubMed/MEDLINE, coupled with the exponential and multitude of data to be captured in each study; due to high-throughput genomic assays, which evaluate hundreds of genes and mutations in larger populations than before.

The problem of tagging all new genomic variants identified in research publications for inherited genetic disorders is a significant barrier to advancing the use of this information to generate clinically relevant knowledge. Among solutions to this is Optra Bio-NLP, a web-based automated annotation system for scientific biomedical English language text. The tool is aimed at identifying genes/variant and disease of interest using a context specific analysis of PubMed abstracts. Optra Bio-NLP will make an intensive use of standard bio-medical ontologies and directories which encode knowledge, to identify relevant entities in several parts of the text. With the use of Technology and Computing Strength, the similarity algorithm would be programmed to create a map of relations/concepts between the bio-medical entities (e.g.: gene/variant/disease/clinical phenotype). If such correlation is found, which our R and D team feels is quite logical, each of the bio-medical entity can be further queried for its ontological structures or to retrieve relevant scientific texts citied in NCBI PubMed. The output of the system would be a robust database of scientific annotations. Information thus generated can auto-populate legacy database with genes, variants, haplotypes and also related diseases or symptoms.

New knowledgebase built with this automated approach will support evidence based variant classification and allow easy interpretation of biologically coherent relationships between clinical phenotypes and causative pathogenic mutations. Further, clinical researchers and geneticist will be equipped with highly significant genotype–phenotype data, to use towards developing new assays, chips and tests for novel disease causing variants for prenatal diagnostics. Finally, in the near future, it may also contribute significantly for generating personalized risk reports for prenatal screening and counseling. Thus automation of this complex process will reduce the work of a clinical researcher, physician and consulting doctor and might eventually prove an efficient tool for clinical applications in personalized medicine. Modern genetic diagnostic companies wanting to develop patented carrier testing platform for inherited disorders can utilize the annotated information to create reference genome assemblies, verify reads of high throughput sequencing and use evidence based variant classification for generating personalized risk reports for prenatal screening and counseling.

Thus using the strength of natural language processing, machine learning and semantics to generate a computerized analysis of possible annotations from scientific abstracts, Optra Bio-NLP will assist in turning information into knowledge by tagging, classifying and discovering new knowledge (Fig. 1).

2.3. Automation of knowledge work: a case study of medical imaging

Another potential area where automation of intelligent knowledge work will prove significant is medical imaging in pathology and disease diagnosis. Immunohistochemistry (IHC) and fluorescent in situ hybridization (FISH) are the two main staining methods used in the pathology. Examination of the microscopic slides is the crucial part in pathology which decides the power of the test. If the microscope slide analysis is made automated it will relieve the human eye and might reduce human errors and will improve the power of pathological tests. Optra Health has developed unique software platforms which can detect, demarcate and annotate 'region of interest', benign versus malignant cells, or abnormal cells in tissue slides, which can lead to unbiased and speedy diagnosis. Using image informatics, these applications can capture image data from various diagnostic instruments like microscopes, tissue slide scanners, Fluorescence and 2D/3D gel scanners, CT, MRI, X-ray and PET scanners. The application can also mine image data from analytical instruments like high content screening, proteomic and functional genomics platform. All images can be annotated with metadata, and analyzed using intelligent algorithms for single or batch processing. Thus, using automation to process clinical, pathological, genomic and proteomic signature data, physicians can integrate all patient specific data to evaluate the disease/condition in a holistic manner, bringing new hope for better and more targeted and personalized treatments, which

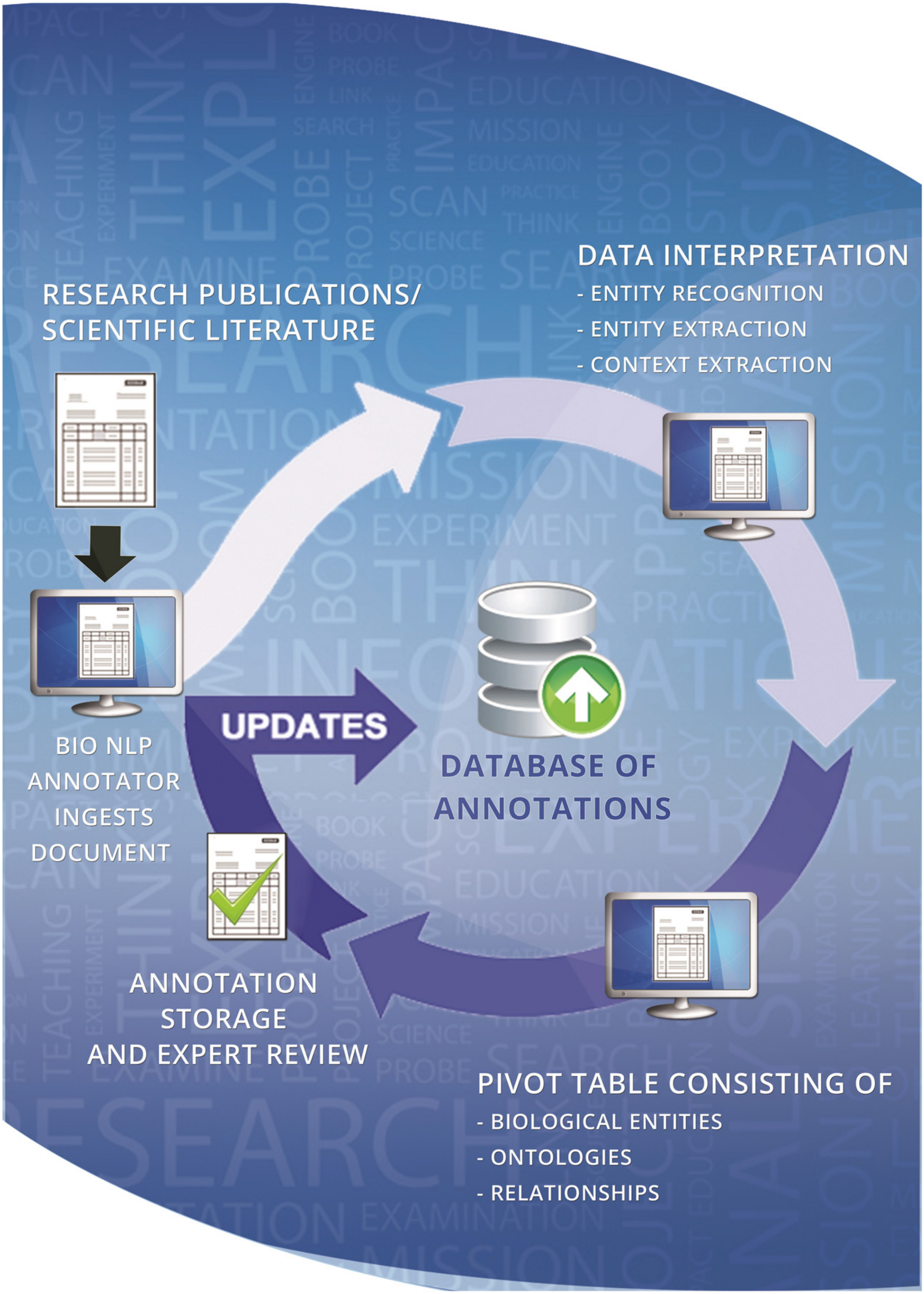

Fig. 1. Optra Bio-NLP workflow.

Table 1
Selected open problems in automation of knowledge work for biomedical application.

Area	Brief problem	Possible solutions
Natural language processing	Handling concepts at different levels of granularity (for e.g.: different nomenclatures of genes, specific and general classes of diseases)	Creating interlinked biomedical ontologies with the use of hierarchical relationships and representations which are easy to mine
Mapping relations and associations	Sophisticated concept mapping and comparing the associations at each level	Developing robust association rule mining algorithms and decision trees using NLP systems to extract relations with co-occurrence statistics
Cooperative learning algorithms	Significant amounts of expensive, manually annotated training data required for machine learning algorithms	Defining a formal model of learning algorithms that can communicate their hypotheses and/or other information in an attempt to greatly reduce the time required to learn
Asynchronous knowledge acquisition	On the fly learning needs to be able to predict the need for extension of knowledge generated by paradigm shifts, as it may be a less frequent occurrence but can lead to considerable changes in the underlying domain	Devising incremental knowledge acquisition protocols based on frequency on past paradigm shift events

may also prove to be significantly cost efficient and timely, in the long run.

3. Challenges ahead

Although the case studies cited above present a proof of hypothesis that many data intensive tasks can be well-formalized by computational approaches, we wish to underscore that the recent progress in this area represents only the beginning of efficient machine learning. There are several challenges that lie before us, to make these automated knowledge workers robust and full-proof. The coming years are likely to see significant progress on the issues in machine learning, artificial neural networks and intelligent programs created to mimic human problem solving skills (Barrat, 2013; Neapolitan and Jiang, 2012; Chen et al., 2013). Table 1 presents a short list of selected open problems and areas for further research in machine-reasoning processes.

3.1. Ethical dilemma: can machines take over decision making?

Finally, some important questions come to mind with automation. Is computational intelligence superior to human intelligence? Can we be certain that computational intelligence will produce error free knowledge? Is it wise for us to rely on computational intelligence and if so is there any limit to our reliance? Can machines take over decision making for which humans rely on meta-cognitive processes, insights and intangible sense of intuition? And most importantly will it be harmless? Other key ethical questions are who ought to decide the use and advancement of AI towards Super AI-intelligence that teaches itself? Will all stakeholders have an equal say in determining what the appropriate development and use of this intelligence is? Ought we to use the full capabilities of computational intelligence, merely because we can? If not, how do we decide how much and when it is appropriate to use?

Eventually, automated knowledge workers may progress to proposing novel answers for evaluation, and some solutions may turn out to be also uniquely successful. However, there is still a long way to go for such computational intelligence to achieve autonomy in decision making. From now till that time, these intelligent software systems will evolve only if they are provided with more and more data, which will improve machine learning, consolidate the neural networks and make them robust for decision making, rather than just operating in decision support role.

Conflict of interest

This article reflects an assessment of the field of knowledge work automation and was authored by Optra Health, one of several companies developing applications in medical machine learning.

References

Barnickel, T., Weston, J., Collobert, R., Mewes, H.W., Stumpflen, V., 2009. Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts. PLoS One 4 (7), e6393 (July 28).

Barrat, J., 2013. Our Final Invention: Artificial Intelligence and the End of the Human Era. Thomas Dunne, St. Martin's Press, New York.

Chen, H., Martin, B., Daimon, C.M., Siddiqui, S., Luttrell, L.M., Maudsley, S., 2013a. Textrous!: extracting semantic textual meaning from gene sets. Apr 30 PLoS One 8 (4), e62665. http://dx.doi.org/10.1371/journal.pone.0062665 (Print 2013).

Chen, X., Shrivastava, A., Gupta, A., 2013b. NEIL: extracting visual knowledge from web data. ICCV.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., Law, M., 2012. Comparison of next-generation sequencing systems. J. Biomed. Biotechnol. 1–11. http://dx.doi.org/10.1155/2012/251364.

Manyika, J., Chui, M., Bughin, J., Dobbs, R., Bisson, P., Marrs, A., May 2013. Disruptive Technologies: Advances That Will Transform Life, Business, and the Global Economy. McKinsey Global Institute. McKinsey & Company.

Mardis, E., 2011. A decade's perspective on DNA sequencing technology. Nature 470, 198–203.

Neapolitan, R., Jiang, X., 2012. Contemporary Artificial Intelligence. Chapman & Hall/CRC 978-1-4398-4469-4.

Quail, M., Smith, Miriam.E., Coupland, P., 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13 (1), 341. http://dx.doi.org/10.1186/1471-2164-13-341.

Toma, I., St. Laurent, G., McCaffrey, T., 2011. Toward knowing the whole human: next-generation sequencing for personalized medicine. Pers. Med. 8 (4), 483–491.

Tucker, T., Marra, M., Friedman, J.M., 2009. Massively parallel sequencing: the next big thing in genetic medicine. Am. J. Hum. Genet. 85 (2), 142–154. http://dx.doi.org/10.1016/j.ajhg.2009.06.022.