semanticweb.com The Voice of Semantic Web Technology and Linked Data Business

Analytics, Health Care / Life Sciences, Scientific and Research Applications

Semantic Tech Turns up Biomarkers and Phenotypes, Avoids Dead Ends and Higher Costs

By Jennifer Zaino on February 26, 2013 10:12 AM

Dr. Carlo Trugenberger, co-founder and Chief Scientific Officer at InfoCodex Semantic Technologies AG, has co-authored a report reflecting the topic he discussed at last fall's London SemTech event: An approach to drug research that relies on identifying relevant biochemical information using the company's autonomous self-organizing semantic engines to text mine large repositories of biomedical research papers.

The model, says Trugenberger, is a departure from many other semantically-engineered approaches to streamlining drug research, which are based on natural language processing (NLP). That's good for extracting information from documents, he says, but not as adept at discovering knowledge. "That's what our InfoCodex software is designed for, to find new facts and hidden correlations" in repositories of unstructured information.

Improving the ability to make those connections is something the pharmaceuticals industry needs badly, he says. "The industry is at a crossroads, in dire straits. Research costs are becoming unbearable, it's more and more difficult to find really useful molecules, and there are more regulatory pressures," he says, which makes it ever more challenging to secure drug approvals from agencies like the FDA. Margins are squeezed and mistakes cost a fortune, so that even large companies can't afford to be led down dead ends. At the same time, the focus is turning from blockbuster drugs to individualized medicine for groups of individuals whose genetic background may expose them to particular conditions, presenting opportunities for those companies that can be nimble when it comes to discovering novel biomarkers and phenotypes.

"Crucial for this development are the biomarkers, the entities – normally molecules like genes or proteins — that are characteristic of particular conditions," Trugenberger says. "Accurate biomarkers have become a huge industry. And finding new biomarkers enables a company to design drugs in a surer way, targeted to certain groups of individuals."

With more and more pharmaceutical data available – big databases of research papers such as PubMed, public clinical trial summaries, internal company information, and so on – in unstructured format, the biological lab process around discovering biomarkers and phenotypes can get a boost from technology. InfoCodex' idea was to leverage text analytics, semantic technology and machine intelligence to read and digest seemingly unrelated papers, clinical trial reports, and other unstructured documents to find correlations among entities, he says. A human researcher on his or her own, after all, can't alone tackle the job of reading hundreds of thousands or even more documents and tracking similarities and correlations among them.

The pilot project that the newly published paper discusses, as did Trugenberger's presentation at SemTech, was done in conjunction with its client pharma giant Merck, using its internal research documents, and Thomson Reuters, using its biomarker databases, as

well as 120,000 PubMed resources, and focused on discovering potential novel biomarkers and phenotypes for diabetes and obesity. The documents themselves weren't concentrated on biomarkers, so it wasn't as easy a job as performing a search across them for keywords like phenotype, gene and diabetes, for instance.

"Some may have contained information that combined with other information in completely different papers pointed to a possible biomarker," he says. "What we were doing is combining information that is seemingly unrelated to biomarkers when taken by itself, but when combined with other information might actually point to a biomarker."

The software combines a very large thesaurus (almost 4 million words and phrases) organised into a seven-level taxonomy of about ten thousand concepts with information-theoretic algorithms and self-organizing neural networks, Trugenberger says. "Information theory is used to transform documents into mathematical models formulated on the linguistic structure, [and] self-organization to semantically organize and match the documents according to meaning. This technology is designed to be able to detect hidden correlations distributed over groups of otherwise seemingly unrelated texts." So, it is not based on analyzing logically a sentence but more on holistic concepts that are distributed over documents.

The results of the pilot identified what he says were several interesting and very valuable candidates for Merck – interesting and valuable enough that Merck isn't allowing these biomarkers and phenotypes to be disclosed. "Each is worth a lot if they are confirmed, which now has to be done in a lab and with human expert assessment," he says. "It is unrealistic to think that a machine with whatever high-quality software can completely substitute for humans. This type of software and machines should be used a pre-process to organize a huge quantity of documents — to do the heavy lifting – and then human experts can then tackle the final step of what is really relevant."

That expertise is needed to assess whether the gene or protein proposed is reasonable and then pass things on to the lab to test. "But," he says, "that is still a huge gain because it's much easier to say yes or no on a proposed list than to establish the list yourself in the first place."

While there was some noise among the results in the first test pilot, Trugenberger says the problem was identified and the technology has been updated to address it. Post-pilot, InfoCodex is interested in conducting projects with other pharma companies as well as it proceeds to commercialize the technology and its process.

In addition to considering selling its services and software direct to companies, it's thinking about doing research in collaboration with medical experts to possibly sell its biomarker discoveries itself. While the first pilot was focused on diabetes and obesity, since Merck's interests lay there, the technology can apply to any other condition or disease, he says – not only around biomarkers but potentially for other molecules as well.

"The idea is to save costs, avoid dead ends and be quicker to the market with new drugs," says Truberger. "So companies can increase margins, avoid dead ends and beat the competition."