Turning Text Into Insights
Interview with Jane Reed, Linguamatics

For podcast release
Thursday, May 18, 2017

KENNEALLY: Epilepsy, leukemia, psoriasis, and dry eye are among the common and the rare diseases that saw critical FDA approvals in 2016 for new treatments. Altogether, 22 drugs cleared US government hurdles last year, less than half as many as in the year before. According to an American Chemical Society-published report, 2017 should see a return to the norm of 30 approvals.

Hello, I'm Christopher Kenneally for Copyright Clearance Center. The scientists who develop these new drugs in the digital age rely on published research for the information that can open a path to discovery. Yet more than 2.5 million peer-reviewed articles appear in scholarly journals in a single year, creating a deluge of data that requires a technology solution.

UK-based Linguamatics is a world leader in deploying innovative natural-language-processing-based text mining for high-value knowledge discovery and decision support. Jane Reed heads the company's life science strategy efforts, working with the pharma, biotech, and healthcare industries to speed up the drug discovery cycle and improve patient outcomes. As an experienced researcher in genetics and genomics, Jane knows the challenges faced from the bench side and the business side when bringing a drug through inception through commercialization. She joins me now from the Linguamatics offices in Cambridge, England. Welcome to the program, Jane Reed.

REED: Hi, Chris. Thank you for the introduction. That's very kind. And I'm looking forward to our discussion today.

KENNEALLY: It's a fascinating area, and I really have to start by asking you, as someone with a background as a researcher yourself – we use the number 2.5 million peer-reviewed articles in a single year. It really is a deluge of data. How are researchers coping?

REED: One of the ways they're coping is – to be honest, they're not coping. I think one of the key challenges is people are making decisions based on subsets of data. Because of that deluge, because of that huge amount of information, people aren't having systematic search or comprehensive search. When you go to Google and you search for a holiday destination or that kind of thing, it doesn't matter if you're not getting comprehensive answers. You don't need it. You just need the top 10 or a few ideas. But when you really want to understand all the genes that people know are involved in a particular disease or pathway, or all the compounds that

inhibit a particular target, or all the different ways that patients are talking about your drug on the marketplace, then just getting the top 10 isn't good enough. You need that comprehensive, that systematic view of the information that's out there.

KENNEALLY: It's a really thoughtful answer, and I guess I'd ask you to tell us a little bit more about that. How does one get that comprehensive view, then, if in fact we're faced with this enormous number of articles and other materials in the published domain?

REED: One of the challenges that people need to overcome is being able to look at the huge variety of data. The often-quoted statistic is that unstructured data makes up the vast majority of information in an organization, maybe as high as 80%. You'll get that in both internal reports and then the external literature or external patterns or social media. And to be able to get a good view, again, you need to take that external knowledge, the external literature, and combine it with your internal knowledge.

Again, one of the key things is things like formats – being able to deal with PDFs, HTML, Word, and being able to make sure you've got the right words to search. It's no good just thinking, well, my favorite gene is (inaudible) or whatever. You need to be able to think of all the different ways people describe any particular disease, any particular gene, in order to again get that comprehensive view.

KENNEALLY: Right. You used the term there unstructured data. For our listeners, we might want to define that. But it's important to have that definition in mind, because in order to do the kind of searching you're speaking about – and we'll talk about text mining specifically in just a moment – it needs to be structured somehow, and a PDF and various other kinds of digital versions of published material can really just be kind of a blob of text. Imagine searching through a blob. It's not very enjoyable.

REED: It's something we do in our daily life. If you go back to that example – if you or I are trying to book a holiday, we might sit down with Google for half an hour and we'll do some searches. You might say I want to go somewhere hot in July, not too hot. And you'll end up writing notes by hand – OK, this is a good place. That's a good place. And you'll end up with a little spreadsheet or a document or some notes where you are summarizing the core bits of information you've been able to pull out of that Google search.

Now, what's Google's doing is it's using your keywords to take you to a whole lot of different web pages, and then what you're doing is you're sitting there and you're reading those web pages. You're reading that unstructured text – all the words, all the sentences, all the paragraphs that somebody else has put in because they're advertising holidays. And your job is to sit down and try and write all that

information down. So you are structuring the content to answer the question where to go on holiday.

Now, that's a really simple question. But when you start saying what genes are involved in breast cancer, suddenly you could type that into Google, and again you'll get 10 hits, 100 hits, 1,000 hits. But imagine then trying to make sure that you can read all the information, pull out the right facts, structure it, and somehow visualize and understand and be able to act on the information you've gleaned. That's a really tough problem.

And historically that's exactly what people have done. All the people listening – when they were doing their research or if they're doing research now, generally people sit down and read free text and create summaries themselves manually. But that means it's not sharable. You can't easily share it across your organization. It's a one-off thing. It's not reusable. It's not comprehensive. And it's not systematic. So you're only ever going to get a subset of the possible information that will enable you to answer that question really well.

KENNEALLY: Right. Thankfully, we have computing power to step in and really transform that activity into something much more comprehensive, better structured, and really ultimately more efficient as a means of coming to an answer. So tell us about Linguamatics and the I2E platform, which as I understand it helps researchers get into text and other content using what's called natural language processing.

REED: Yeah, thanks. What natural language processing does is it provides a toolbox to enable a computer to really understand how people have written things. I2E, our solution, has a combination of tools that involve both linguistic tools – so that means it breaks down every particular sentence and understands what the subjects are, what the objects are, what the relationships between those, and then a set of tools around semantics or vocabularies or ontologies. That means that using text mining, you can build a search to say, OK, I want all the lists of words around a particular set of genes, and I want all the lists of words around a set of relationships – let's say verbs, so is involved in, is a component of, has a role in. And then a set of words around your diseases – maybe breast cancer again. Let's stick with breast cancer.

That means that the tool will understand the difference between a gene being involved in breast cancer or not playing a role in breast cancer or may have a role in breast cancer. If you're doing research, it's really important to be able to pull out those different relationships between any of the key players – the genes, the tissues, the compounds, the drugs, the diseases – and understand those relationships. Is it involved? Is it causal – is it being caused or is it not causing a particular change?

KENNEALLY: Right. This kind of text mining that Linguamatics makes possible – it's not new entirely, but it really is accelerating in its adoption, particularly within the big pharma. And Copyright Clearance Center and Linguamatics have recently partnered to allow for a truly comprehensive search of published text. You spoke about how important being comprehensive is. Because the search that we allow for goes beyond abstracts and gets into full text. Why is full text important?

REED: One of the things is many people find that pulling out information from scientific abstracts is invaluable. It gives you a high-quality summary for your research. But many facts and observations are excluded from the abstract and only show or only are described in the body of the full text. In particular, a lot of our customers find there are often key facts provided in tables. And as anyone who's tried knows, extracting and summarizing tabular information from tens or hundreds of papers is hugely valuable, but without text mining, very time-consuming.

For example, one of our customers, Shire, were looking for mutations around a particular rare disease gene in order to enable better patient treatment. They found that a lot of those mutations weren't described in the abstracts, but were in the tables within full-text papers. So searching the abstracts again might lead you to the right paper to read, but it wouldn't provide you with the full view you need.

KENNEALLY: It's really interesting. This kind of comprehensive search – it's useful in the development of a drug or sort of leading down that path of discovery that we spoke about in all kinds of phases, from the very beginning of research, the early-phase research, through pre-clinical drug safety tests and drug repositioning, and even for competitive intelligence. So the kind of activity we're talking about – it may begin at a lab bench, but it sort of works its way through the entire workflow. Can you tell us a little bit more about that?

REED: In many ways, if you think about the whole drug discovery, drug development process from the whole bench to bedside, you need insight at every single stage to answer questions. Early on in target discovery, you might be interested in searching all the literature for specific genes involved in a key therapeutic area, or you might want to search patent literature to understand the landscape around a particular assay technology, for example. Right from target selection or lead selection, pre-clinical safety, clinical safety, clinical trial design, and forming your protocols, even into post-market around being able to get a good view of patient-reported outcomes, a lot of that information will be in free text, and a lot of that information can be extracted if you use text mining tools.

KENNEALLY: All right. We are speaking right now with Jane Reed, who heads life science strategy efforts for UK-based Linguamatics and learning about text mining and why it's so important to the development of the next wonder drug. Jane, you mentioned a particular customer that you work with, Shire, but I'm sure there are many others around the world in the healthcare industry and pharma as well. These

kinds of efforts – this text mining that Linguamatics enables and Copyright Clearance Center works with you on – it has a number of benefits. It can reduce cost, I would imagine. It also, it would seem to me, focuses people's time and efforts, and that's really critical. All those hours at the lab, they have to amount to something very quickly, because many of these companies are really – the development of a new drug can be years. Anything that shortens that development span is probably a very welcome addition to their technology toolbox.

REED: Yes, absolutely. Again, if you think about that whole drug discovery development pathway, there are so many questions. Ensuring that you can maintain a competitive edge and really get the right answers fast or backed by comprehensive research, then any individual customer can make gains all the way along that pathway. As you say, we have pharmaceutical customers using text mining across full-text literature to speed up target selection, pre-clinical safety review, and all those little incremental gains at each step – you can get a significant reduction in the overall time, and hence cost, of bringing a drug to market.

KENNEALLY: Indeed. For those who are listening who may not be working in the particular industries that are the targets here, they know about search. One of the things that we always try to clear up around all of this is when we're talking about text mining, this is not search. Even search on Google Scholar, a kind of Google on steroids, is nothing compared to the kind of activity that you're describing.

REED: Yes, you're absolutely right. Everyone is so familiar with search. It's amazing. When you think about how the landscape of search has changed in the last five years, 10 years, 20 years, people are so used to being able to find information at their fingertips. But search only gets you so far. It gets you to a set of documents. It gets you to the corpus of information, the body of information that you want to summarize. And maybe it pulls out – a lot of search tools will highlight the keywords. But in order to really start being able to feed that into analytical packages, in order to visualize it – because as humans, we process information visually much better.

So if you want to be able to graph it or chart it or dashboard it, or if you want to feed it into – you hear a lot about artificial intelligence and machine learning. If you want to extract the right set of facts to feed into your machine learning, you need not just to find those or search those facts, you need to extract them. And you need to extract them with the right context and in the right order and with the right vocabularies in order to make sure that you've pulled out the relevant facts in the right relationship and structure those.

I think that's something you said right at the start, Chris. People are so used to being able to analyze structured information. Everyone is familiar with using Excel and tools like Spotfire or Tableau or those kind of graphical tools to take structured data and be able to quickly analyze that. The core value that text mining brings is

that you can turn all of your unstructured knowledge into structured information to visualize it, to analyze it, to create those actionable insights.

KENNEALLY: Indeed. That's the way, I think, to close, Jane Reed, is that we're speaking about here the effort to turn texts – published material of all kind – into insight. It really will have dramatic results for all of us living on Earth, because the kinds of drugs that will emerge will help tackle diseases that we once thought were uncurable.

REED: Yes. The company I mentioned before, Shire, they're working in rare diseases. Even across the common diseases – the big-board metabolic diseases, cardiovascular diseases, so many different types of cancer, neurodegenerative diseases – every day, there's new research being published. If you want to stay at the forefront of that, you either rely on curated databases – there are many of those. But in order to get the most up-to-date information to be kept alert, alerted by that, to be able to pull out the core information that you need, then using a text mining tool, ideally in conjunction with something like Copyright Clearance Center's tool for rapid access to full-text literature, that really helps you move your research forwards.

KENNEALLY: Again, thank you, Jane Reed. Jane Reed is head of life science strategy efforts for UK-based Linguamatics. She's speaking with us from her office in Cambridge, England. Jane Reed, thanks for joining us.

REED: Thanks very much, Chris. It's been a great conversation.

KENNEALLY: Copyright Clearance Center will be among 3,300 life science, pharmaceutical, clinical, healthcare, and IT professionals from more than 40 countries at the Bio-IT World Conference and Expo on May 23-25 at the Seaport World Trade Center in Boston, Massachusetts. We invite you to visit Copyright Clearance Center at our booth, number 548, to talk about your R&D team's information challenges and to learn how CCC's solutions can help.

This year's conference features over 200 technology and scientific presentations covering big data, smart data, cloud computing, and trends in infrastructure, new technologies, and high-performance computing. So you won't want to miss it. Again, you'll be very welcome to drop by our Copyright Clearance Center booth at Bio-IT World, booth number 548. We'll be in Boston, our hometown, at the Seaport World Trade Center May 23-25.

For all of us at Copyright Clearance Center, I'm Christopher Kenneally. Thanks for listening.

END OF FILE