KENNEALLY: How quickly do you read? According to results of an online speed reading test by Staples, the office supplies company, the average senior executive reads 575 words per minutes, while the average college professor clocks in at 675. The rest of us manage only less than half that volume, about 300 words per minute.

Why does this matter? According to the U.K.-based Jisc, the global research community generates over 1.5 million new scholarly articles each year. Even a champion speed-reader would never be able to keep up or ever do anything else but read. That’s why life science companies increasingly are relying on text and data mining software to identify and extract from this information mountain the research equivalents of diamonds and rubies and sapphires.

At Copyright Clearance Center, we work with life science companies around the world to enable access and sharing of the tremendous volume of published research available today. We understand what a daunting challenge it can be to manage and mine this information mountain range. Within the millions of words and illustrations in your libraries and databases, the potential for insights and innovation is enormous, for drug discovery and clinical trial development to drug safety monitoring and competitive intelligence.

For those of you who are not already familiar with us, Copyright Clearance Center is a global rights licensing technology organization. We provide content workflow and copyright compliance solutions for more than 35,000 companies with employees located around the world. CCC also offers extensive copyright education and training programs from webinars like this one to online videos.

CCC is developing for availability later this year a service that provides full text content for mining from multiple publishers in a single format. The work we’ve undertaken on
this effort has led us to speak with the leading publishers and researchers in the field, including our guest today, Lars Juhl Jensen, professor at the Novo Nordisk Foundation Center for protein research at the Panum Institute, Copenhagen, Denmark. Welcome to our program. Professor, Jensen, welcome.

JENSEN: Thank you very much.

KENNEALLY: We’ll tell people briefly about your background, Professor Jensen.

Lars Juhl Jensen started his research career in Søren Brunak’s group at the Technical University of Denmark. In 2002, he earned a PhD degree in bioinformatics for his work on non-homology-based protein function prediction. During this time, he also developed methods for visualization of microbial genomes, pattern recognition in promoter regions, and microarray analysis. From 2003 to 2008, Professor Jensen was at the European Molecular Biology Laboratory where he worked on literature mining, integration of large-scale experimental datasets, and analysis of biological interaction networks. Since the beginning of 2009, he has continued this line of research as a professor at the Novo Nordisk Foundation Center for Protein Research at the Panum Institute in Copenhagen.

He is also co-founder and scientific advisor of Intomics A/S. Professor Jensen is a co-author of more than 100 scientific publications that have in total received more than 10,000 citations. He was awarded the Lundbeck Foundation Talent Prize in 2003, and his work on cell cycle research was named Breakthrough of the Year in 2006 by the magazine Ingeniøren.

In addition, his work on text mining won first prize in the Elsevier Grand Challenge: Knowledge Enhancement in the Life Sciences in 2009, and in 2010 he was awarded the Lundbeck Foundation Prize for Young Scientists.

Professor, Jensen, this is still very early days in text mining. I wonder if you could give us a sense of how far we’ve come. Clearly, we’re not at ground zero. There’s been a great deal of development involved, but still a long way to go. Are there certain areas that have come along further than others? For example, identifying names. How well can you do that contrasted, say, with really understanding relationships within texts?

JENSEN: I would say certainly something where we have the best performance so far, some of the most studied topics, are certainly the things related to identifying names in text.

Unsurprisingly, although you can of course gain something from looking at the context around names and so on, really what everybody realizes is that the key to recognizing names in text is to have a good dictionary. If you have a lot of bad names in your
dictionary, you’re going to make a lot of errors. If you have a dictionary that is very incomplete so there are lots of synonyms for each thing that you don’t know, you’re not going to do a good job.

But I would say, especially for something like recognizing gene and protein names in text, I think the methods are beginning to converge when you look at the assessments or informal competitions that are within the area. Things seem to be sort of converging. The best groups are doing roughly equally well, and the performance is not increasing dramatically every year anymore.

So that’s getting to the point where the performance is really useful. It’s at the level where of course it still makes errors, but then again, if you ask two human beings to go through the same text and mark it up, they will also not fully agree and they will also make mistakes. So one should not hold the text mining up against some extreme standard that human beings is (sic) also not capable of doing.

KENNEALLY: Indeed, and as I say, I think that we recognize here at Copyright Clearance Center, as you must in your research lab, that these are still very early days when it comes to text mining. But again, I’m trying to get a sense from you to –

JENSEN: Text mining has been going on for decades in terms of research, but it’s more been people doing text mining because they were interested in that as an intellectual exercise, so to speak. It was the challenge of how do you do text mining.

What is early days is more the people like me who do it from a very applied standpoint that I’m not doing text mining because I’m interested in doing text mining research. I’m doing text mining because it’s come to the point where the tools are actually useful for solving real problems.

KENNEALLY: A very important distinction, clearly. Helping to make those tools more useful, of course, is really what this presentation is all about. So, a question I want to ask you about is, give us your own personal assessment of how the access to information, how – I was going to say, how open the access is, but that sounds like open access. I suppose open access has made a difference in what is available, but what else are you hoping to see?

JENSEN: What I hope to see is, of course, some of the things that you are also working on in terms of being able to actually not have to run around to each publisher individually to obtain the text. It would be useful to have places where you could go and get hold of the text from one place in one format and others at – preferably with a unified license so that you actually know that you’re allowed to do the same things with all the text.
Because it’s very difficult to do that kind of micromanagement where you have to deal with the different papers from different publishers. You need to technically obtain them in different ways. You have different legal agreements with all of them. You’re not allowed to do the same things with all of them. That is really – you don’t have time to do that as a researcher, and that’s why nobody’s doing it.

KENNEALLY: Right. Well, Professor, Dave wants to know if there’s a specific example you can share where text mining has made a surprising breakthrough, has provided us with a discovery. You made the point that text mining isn’t news, so probably it’s not so much something that happened yesterday, but really an example of where digging in and doing this type of research around looking for relationships in text and seeing common terms and so forth –

JENSEN: It really depends on – when it comes to doing discoveries as such, I would say then of course, looking at the biomedical literature is not the place to do text mining because you would argue if you’re doing text mining on the biomedical literature and you’re doing it right and you’re not making any mistakes, then you’re only going to discover the stuff that is already published.

Of course, the thing is, without text mining, people are very likely to overlook a lot of things that have been published, so it’s still a very useful tool. It’s very difficult to see how you should do text mining on the published literature and actually discover something new from doing that.

I think the places where you have chances to really discover something new by doing text mining is more in the area of work like I discussed with mining electronic health records, for example, because that allows you to collect a data set on, for example, how many of the people who get a certain drug get a certain adverse drug reaction, and that way, you follow up on things and discover things that are simply not known to science because different doctors see different patients, and until somebody goes through all the text written by all the doctors and put it together in one big statistical analysis, you’re not going to find that a certain drug presumably causes a certain adverse drug reaction.

So it’s more those kinds of areas where you’re using the text to gather a lot of information in one place in a database so that you can do statistics on it afterwards to discover something that is a trend in the data but which was not actually written anywhere.

KENNEALLY: So if I hear you right, it’s really about leading you to ask more questions. A point that Kim on the webinar is making is about causality, and I don’t know life sciences, but I know that co-relation is not causation, right? So what you’re suggesting in regards to
reactions to drugs is there may be some data pointing you in a direction. That’s where the research begins. It doesn’t end at that point.

JENSEN: For sure, for sure. I mean, firstly, of course whenever I do text mining, if you think you’ve discovered something by text mining, the first thing you do is you go in and you look at your data. You look at the text mining results to make sure there is not any error.

But on top of that, what we are able to do – I didn’t talk about the details of how we do the pharmacovigilance study. What we do there is we have time stamps on everything so we know when the patient started getting a certain drug, and we of course also have time stamps on all the notes about each patient. So for that reason, we are able to do some sort of causation analysis based at least on temporal correlations, that you can see that it’s patients that start getting a certain drug that will in some period of time afterwards start getting a certain adverse drug reaction. In that sense, you are able to do something that is causation.

Again, it’s hard to know. Things can always be in direct – also, patients often get multiple drugs at the same time, so indeed, it’s more something that points you in a direction of this looks suspicious. There seems to be too many people who get this side effect after getting this drug. Maybe we should look into that.

KENNEALLY: We have a question from Rich regarding whether or not you feel text mining should be what he refers to as an iterative process. In other words, not just one pass, but several passes are necessary to really uncover what a scientist or researcher is after.

JENSEN: There are sort of different ways of doing it, and that perhaps also leads a little bit into the difference between what the system you’re developing is currently set up to do and the kind of approach I’m taking, because indeed, what some people do is they have this more iterative approach where they start by doing some queries that somehow nail down what is hopefully the relevant subset of papers and then subsequently do the text mining for gene names and correlations and so on in that subset of papers.

Whereas the approach I take is the shotgun approach where I say, OK. I just take all the text and I do the whole thing in one go. I don’t think there’s a disadvantage to doing the whole thing in one go other than the fact that of course it’s computationally rather intensive because processing – you’re putting all the text through the full pipeline.

It depends on what you’re trying to do. If you’re interested in, for example, coming up with a gene network for some specific pathway that is relevant for studying some disease or for drug development, then taking this approach where you try to first make it on the
papers by doing searches and then afterwards doing text mining on the papers you obtain makes perfect sense.

If you’re trying to make general databases where people would come and search for anything, like the STRING database where people might go and search for any gene, then you can’t really take that approach because you would have to do so many different queries that I don’t see how you would do that. So then it makes more sense to just say since we want to make a network for all the genes, we just take all the literature and do the whole thing in one go.

KENNEALLY: All right, Professor. We have time for one more question, and it is the big question from Steta. And Steta is asking what will be the next big thing in the world of biomedical text mining? So if you’ve got a crystal ball there at your research desk, let us know what’s the next big thing here? What would you like to see next in biomedical text mining?

JENSEN: Oh, that one is of course always difficult. There are various fronts where there’s certainly things happening. Like, for many years, it’s been really pretty problematic having a computer trying to parse text, parsing it as in actually understanding the grammar and pulling out relationships that way rather than just saying this sentence mentions this and that. They probably have something to do with each other.

There’s a lot of progress being done in that field in terms of the precision getting up there where it’s useful. The problem so far on that is largely speed, that the best methods of computation are very intensive. But although the literature is growing fast, thankfully compute power is growing faster, so I think we are getting there where the best methods are good enough but they’re actually getting useful. And at the same time, we’re hopefully getting enough compute power to be able to actually put all the text through it.

On top of that of course, also going to the point of using all the text. And there’s a lot of challenges there because there’s of course the abstract versus full text paper. But it’s also getting, as everybody knows, more and more papers have supplementary material that is often far larger than the paper itself, so also getting to that level.

There are people working on, say, when you’re looking at mining articles, it’s not just the text. You also have things like tables. How can you have a computer automatically understand the structure of a table and pull out the data from that?

There are lots of things happening in those different directions.
KENNEALLY: It’s a fascinating topic and one rich in potential. We certainly appreciate your presentation today and your very candid, thoughtful answers to those questions from our audience. I want to thank you, Professor Lars Juhl Jensen at the Novo Nordisk Foundation Center. Thank you so much for joining us.

JENSEN: Thank you, and thanks everybody who joined.

KENNEALLY: For all of us at Copyright Clearance Center, my name is Chris Kenneally. Thanks so much for joining us.

END OF PODCAST