Intervju med Sonja Aits, Lunds universitet. 2020-11-18

Intervjun ingår i dokumentationsprojektet ”Digitala innovationer i skuggan av Corona”. Intervjuledare är Tekniska museets intendent Peter Du Rietz och intervjun genomfördes via videosamtalssystemet Skype. Både intervjuledaren och informanten genomförde intervjun från sina respektive hem. Under intervjuns gång inträffade ett antal tekniska missöden, vilket gjorde att inspelningen fick pausas. Detta transkript bygger på en sammanställd, hopklippt film.

Transkriptet är utfört av företaget Rappa tag och har därefter redigerats av Peter Du Rietz.

I:	Intervjuare (Peter Du Rietz)
SA:	Sonja Aits
[inaudible]=	Hör inte flera ord eller mening/ar
[name]=	Uppfattar inte vilket namn som sägs
?=	Hör inte enstaka ord
…=	Paus, avslutar inte meningen, blir avbruten
[utskriftskommentar]=	Rappa Tags kommentar
Gulmarkerat=	Kontrollera ord/mening

Transkript

I: So my name is Peter du Rietz, I am a curator at the National Museum of Science and Technology, Tekniska museet. Today's date is November eighteenth 2020 and this interview is a part of the project Digital innovation i skuggan av corona, Digital innovation in the shadow of the corona crisis. And I have with me Sonja Aits from Lund University. Could you please tell me a little bit more about yourself?

SA: OK, yes. I am Sonja Aits and I am a research group leader at the faculty of medicine at Lund University. I am a cell biologist by training, so I have a PhD in biomedicine but after that I changed track a little bit and now me mostly do computational research involving artificial intelligence. As you can maybe hear I am originally from Germany but I have now lived in Sweden since 2003 except for a small stint in Australia for two years.

I: Okay. So how did you get into this field.

SA: I guess I have always been interested in natural science per se and that is what attracted me to cell biology to start with. And what attracted me to the field of artificial intelligence is the fact that I was just producing so much data in my lab work that I could not handle it anymore. So I knew I had more information than I could analyze myself, so I was looking for new tools to make sense of the massive mountains of data I was generating.

I: Okay. So you are interested in computers and IT personally and privately.

SA: Yeah, you know... I have never been a gamer or something like this in my free time to an extensive level. I have played a couple of computer games, I have always enjoyed it but somehow never gone into it massively. So I only started programming as an adult.

I: When where you born.

SA: I am born in Münster in Germany, that is northwestern Germany but I never lived there. I immediately moved to Bavaria to a city called Fürth. So I am grown up there.

I: When was your birth year, if I may ask.

SA: 1981.

I: Okay. So now we have a little bit information about you. Now I was planning on going forward to talk about the project. So can you please tell me a little bit about this project Artificial intelligence-based knowledge curation to direct COVID-19 research and public health efforts.

SA: Yes. So the main part of this project is that we are using artificial intelligence-based text mining tools or rather developing them in order to quickly process texts that are related to coronavirus research so that other researchers, for example virologists, can quickly extract the information they need for their drug development or vaccine development. It could also be used then by other people who want to quickly gain an overview over texts. And these texts could be either the form av scientific literature or it could also be patient journals.

I: And how did you come up with this project? What is the initial ideas?

SA: So originally we have done a lot of text mining research in connection to finding out more about cell death and lysosome, which is my cell biology primary interests to start with. And there are hundreds of thousands articles written about cell death and I just felt that it was impossible for me as a scientist to get an overview over this amount of literature. But at the same time that there could be pieces of information scattered across all of these articles that would... if we can put them together... give us a comprehensive image of cell death. So this is how this project started and then when the pandemic came along, we really wanted to help in our group as well so we figured, what is the thing that we can do. So we looked at the tools we already have and we basically switched the target. So now, instead of trying to find out more about cell death, we are trying to find out more about the covid-19 and the coronavirus itself.

I: So you had a lot of the software really in place before the coronavirus came along.

SA: To some extent, yes. But it was not a complete pipeline that we had so we still need to develop it further and we also need to adapt it to covid-19 and the additional challenge which... we had never worked with Swedish texts before so all the work we had done was adapted to English texts. But in the context of the covid-19 we also thought it was very important that we get a tool that can analyze Swedish texts, especially if we want to extract information from patient journals, these are of course in our country written in Swedish. And this was completely new for us so we kind of... we had the underlying strategy and we had the underlying ideas of how you can build these tools, but we had not built the tools that we could use right out of the box so to say. So this is what we are trying to do now, we are trying to modify the tools that we have in a way that they also work with Swedish language.

I: So the software is scanning texts in English and in Swedish, any other languages.

SA: At the moment we are focusing on two which is hard enough. But in principle the same technology can of course be applied to any type of language. So even languages that use a different type of alphabet. But at the moment we are sticking with two... I also think it is important that for natural language processing research, which is the type of artificial intelligence that we are using, it helps if you understand the language that you are working with and the reason for that is that... if you are developing, in the beginning your system, your model makes a lot of mistakes and you cannot understand the system behind the mistakes if you do not understand the language. I mean, you have true and false and the computer can see whether the result was true and false. But if you want to get a deeper insight, you need to be able to verify this and to also look at the fine details. For example, does it always pick up a special word which is similar in that language, but if you do not understand the language then these are the types of mistakes that you may not be able to analyze. So that is also a reason why we are sticking to Swedish and English because both... I myself but also the people in my group, we understand Swedish and English, but French or Spanish or so would be harder for us.

I: And the text input, is that just or mainly scientific articles or is it other kind of materials?

SA: So the tools, when they are complete, they will in principle work with any type of text. But we are focusing our evaluations at the moment, and the finetuning so to say, to two types of texts. One is

scientific literature and that is in English, so all the scientific literature written today or at least most of it is written in English. So even if it is produced in France or in Germany or in Sweden or in Denmark, it does not matter, everyone writes in English. So this is why we are focusing on English literature and the reason we are focusing on literature per se is because the most information about the coronavirus is hidden in this type of texts. So we can also analyze newspaper articles or social media feeds or stuff like that, but if we want to guide the science we should look where the science is written and that is of course in the primary source in the scientific literature. But in addition to that we also developing a second type of text and that is patient journals, so every patient journal that is written about a patient in the hospital today has a number of different fields which is like name, date of birth and so on and also lab results potentially. But there is also always a part that is free text and if you want to do an evaluation for example to try and find out which patients have a higher risk of dying from covid-19, then it helps if you can also analyze this free text because this may describe where the doctor may write, okay... patient has a history of anemia or patients mother has diabetes and so has several other family members. So this kind of information is otherwise not accessible to systematic study or you would have dozens of medical students having to read through them and then kind of fill this over into some kind of Excel-sheet or another type of structure database. So in order to rapidly analyze this information form patient journals, we are also developing tools for this type of text and that is primarily in Swedish.

I: And the patient journals, are those available? How... I mean...

SA: No. Not just like this, no. As for any kind of clinical research that involves a patient or patient material, first you have to apply for an ethics permit. So you have to describe exactly what you are trying to do with your project and what type of data you need and how you're going to handle it. And then you have to submit that to the ethics review and this has been approved for our study. And then afterwards, the next step is to get actual access to the data and this is, in our case, provided by the hospitals of Lund and Malmö for the first study that we are doing. And we have a collaboration with a doctor in the emergency room in Malmö who initiated this project with us. So you cannot just find this type of data on the internet or even in some kind of researcher database. This is highly sensitive material and you also, you know... you cannot send it from one researcher to the other by email and stuff like this. There is a lot of restrictions on patient journals, and that is good. Because they are, in a way, the most sensitive type of information we have about people because everything can be in a patient journal. Everything from a complete patient history to information about their family to information about their mental health and all these kind of information and it should be treated respectfully and only be used in a way that is absolutely necessary for research.

I: Will it get anonymized in this or is it as it is, so to speak?

SA: Yes, it will be... I mean, we are not interested in who has a specific disease, because we are not trying to extract information about a single patient so in this project we are trying to gain a general overview over the response and the symptoms of coronavirus patients. So for us, the individual is not important and this is why yes, we will anonymize. And the primary reason also is that for us it will make it easier to work with the data because there is less restrictions on it if it is anonymous.

I: I have been talking with quite a few researchers now and several of them has been talking about, getting access to this kind of material has been a major obstacle. What has your experience been?

SA: Well, actually the clinical data... I cannot tell you yet because we have not received that yet. The ethical permit has only just recently come through, so now we have applied for data access but we have not gotten the data yet. So you can ask me again in a couple of weeks... or months, hopefully not... how fast we gotten the data. But I know that this is a major bottleneck for many, many type of clinical studies whenever you need data, whether it is patient journals or other type of data. And I think, this is not that the hospitals are unwilling to give us the data, not at all. It is simply that they do not have the resources and dedicated staff who does nothing but deliver data to researchers. I mean, they are hospitals, they have to use all of their resources, especially right now, to treat patients. So you know, researcher is important but for hospital it has to come second. So this, I think, is something that maybe, over time, will change that you know, every hospital will have people that are properly managing this type of data and then also can quickly provide the data. But right now this is not the case, at least not in the hospitals that we work with. I mean, they are trying their best for sure. That I have no doubt.

I: And the research articles. How available are those?

SA: That is much easier.

I: But are they on public domain or...?

SA: Some of them are. So there is, for medical research, a primary database... or rather two primary databases... that collect more or less any kind of medical article or even biological article or biochemical article that is published, with a short summary which we call an abstract which is often maybe around two-hundred words long summarizing the findings of the article. And this is available for almost all articles that are written and these are freely available. So everyone could just go to these databases and look for them. Even you could or... you do not have to be a scientist and you do not even have to register to read that. So this is freely available. Then if you want to have access to the full text scientific article, it depends. If it was published with so called open access, which means that the research team paid so that everyone could read it, then it is freely available in the same databases. It is just a click away. I would say it is about a sixth of the articles has that normally. I think it is roughly five million out of the thirty million in the database that are available in this way. But other articles, then as a researcher you may still have access because your institution may have a subscription... say to Nature or Science, Cell and so on, these important journals... so then I can still click and get access but maybe you could not unless you paid for the article, then of course you could also get it. But in the context of this pandemic, most journals and publishers have actually made all the coronavirus related literature freely available. So even though they might normally be behind a paywall and maybe my institution did not have access, I would not have trouble accessing it. In this case actually they went out and changed their policies and they said, this is too urgent and here is all of coronavirus literature, everyone can access it. So there has been a major change in this context which is very good. So there has also then been an American institute which has a compiled a large collection of coronavirus articles already in the same format. Because other ways... the articles come in different formats and that can also hinder computer science research. So they have made the work to kind of standardize this in a way and make it available as a collection that can just be downloaded basically with, you know... just from one point source to speed up the download and so on. And this comprises not just the covid-19 articles but also articles about other coronaviruses. Because often we may find clues as to how it should treat this coronavirus by understanding how we could target other

coronaviruses, because they are related of course. I mean, this one is called SARS covid-2 and that is because there is already a number one and that is from the SARS pandemic. So I mean, we can learn from this and this is also included in the collection. So I think it is over three hundred thousand articles at the moment, they update it continuously.

I: Could the covid-19 crisis mean a breakthrough for public access?

SA: Yes, well maybe. Because on top of what I have just described, many researchers right now are actually not waiting until their material is published in a journal, but they are releasing it freely before that in so called pre-prints. So in addition to the databases I just described, there are databases where basically researchers can upload a manuscript. The problem with that can be that no other researcher has gone through the material and given feedback for improvements, which is normally part of the publishing process, that you submit, you get feedback, you improve, you submit again... and this can even go through several rounds which we call peer review. So now, this is not happening... people first put it out there but then then still go through the peer review, but that is only afterwards so they make it available already as a manuscript. And I think this has changed in the medical field. This has been common practice in other fields already, for example computer science, even before this pandemic. And it was starting kind of to pick up in medicine as well in the years before, but now this has really gotten a boost. So the manuscripts that are published, they may not always be in the final shape but at least... often they include the raw data and so on. For people who are not trained as scientists, maybe sometimes they will not see weaknesses in articles, but as another scientist you often can spot them. I mean, the people who review for medical journals, they are also just ordinary scientists. So of course, if I am a specialist in virology and I read a manuscript about virology, I can spot if there were flaws in the data or if there were some issues. So it is more a problem that these pre-prints maybe then get reported on by journalists who do not have the competence to always judge the weaknesses of an article. But I think, for general science, it is a huge advantage because now we just have access to information much faster and maybe one scientist can read something and then already build their own work based on this new information much faster than in other cases. A peer review process can take months, so now this waiting time has been taken away. And I think many scientists have seen the advantage of that so well, yeah... maybe it will change now that this becomes even more prevalent. I hope so. We are putting all of our science out as soon as we develop it.

I: Might be something like a medical Wikipedia or something like that.

SA: Actually, much of Wikipedia is... the parts that are about science or proteins and so on... actually written by scientists because Wikipedia has a lot of information for example about specific genes and the only person who has an interest in writing a Wikipedia article about the specific gene is either someone who has a gene defect maybe in this gene or has another personal connection to this gene and that is why they put the information there, or it is actually a scientist working on this gene who wants to share the information about their favorite gene or their favorite protein or their favorite...

[audio cuts out]

I: I think I lost you there for a little bit. Are you there?

SA: I am here, yes.

I: OK, I hear you but I can't see you.

SA: This is the wonderful world of remote interviewing. I have recently been at an event where we had a panel discussion, but halfway through the panel discussion the panel got disconnected from the rest of event. So we could still talk to each other but not to the rest of the audience.

I: How about medical datasets, are you scanning databases with medical datasets that are open source?

SA: At the moment we do not do systematic scanning of it on site of the text mining project. But we are collecting datasets that we can find, both in our own GitHub-page and also, Sweden has developed a data portal for covid-19 data which is called Swedish covid-19 data portal and this is also part of the covid-19 research program that we are part of. So my group, whenever we find interesting data or other resources and tools, we kind of connect that to the data portal so that it is out there. So basically, I am curating the resource section of this data portal which includes for example artificial intelligence models but also infrastructures that are useful for scientists and other types of datasets that we find. And in addition, this data portal is also connected to the big bioinformatic databases that we already have around the world. For example with proteomic data, genomic data and so on and also to the European sister site for it. So yeah, we are collecting data but we are not automatically crawling the web to find datasets even though this is something that potentially also be done. Especially with these text mining tools because the same way you can analyze literature or a medical journal, of course you can also analyze the texts on a website with text mining tools. So we could potentially afterwards try and tweak these to kind of find mentions of covid-19 datasets on the web. But this is not something done in a second so at the moment we are not focusing on that.

I: Okay. So is it the SciLifeLab initiative that you are talking about?

SA: Yes it is. So its sixty-seven groups in total I think. We also have meetings with each other... which is also different from kind of before, that they are trying to connect us better so that we can exchange information and this has already led to the fact that, for example, we are now helping another group in this program... both with text mining but also with image analysis, which was not our original application to this program. But we also have competence in artificial intelligence for image analysis. So in the discussions this came up and so now we are also supporting one of the groups with that.

I: Yeah, I know there are a few projects in this program that are working with image analysis. So you are helping out each other in that?

SA: Yes. I think this is the best way to do the science. I know that some people have a very competitive approach to science but my group does not have this approach and actually many other scientists also do not have this approach. It is very common that we collaborate and that we do help each other as much as we can, because in the end, I mean... most of us are in science because we want to make discovery faster and we want to improve in the medical field. We want to improve the wellbeing of the society and if I can develop a text mining tool that delivers results that help another scientist set up a drug screen that finds a drug to cure covid-19, this will be a huge success. And then it does not have to be me who finds that drug. None of us has the competence to do all of the steps

from beginning to end. So it is much better to collaborate and also more fun, otherwise it is very lonely if you just try to do everything by yourself. You get better ideas by discussion with others and sometimes... it is the same way as any other field, if you talk to people with diverse backgrounds and you kind of put ideas back and forth, then often you come up with the best stuff.

I: Could you please describe how you are using this tool. Is it like a sharpened Google search or what is it like? How do you work with it?

SA: I can give you one example. So one very clear example is that we are trying to identify all drugs that have a potential link to coronavirus. So drugs that could potentially be used to treat this disease. So what we are doing is we are scanning the literature for mentions of drugs or chemical names... drugs are chemicals... so we are using the system to scan and produce basically... first of all kind of tags in the text wherever we find, like it would highlight in the text wherever there is a chemical. But then it can also produce a long list, kind of summarizing all the chemicals we can find, then we can... as a next step... compare these chemicals, for example to drugs that are already approved for others diseases or drugs that are known, drugs that are in clinical trials. And then the next step could be to actually test whether these drugs work in coronavirus. Because, from the text we can also extract relations, something like... this activates that or this drug inhibits that. And we can do that basically, the two pieces which are connected we can exchange so either drug or a protein, we could also find a protein that is connected to coronavirus and this way we could maybe find something about the genetic profile connected to risk factors for example. So this is what we are doing now and in the end we want to kind of connect this to a network of relations. So if A connects to B and B connects to C and C connects do D and D connects to coronavirus, we also want to find that entire chain so to say. And often this chain is of course not described in a single article and this is the advantage of using the AI. A human reader does not have the possibility to keep all of the stepwise connections in the head because it is just too much information to memorize. So this is kind of what we are trying to find. We have pieces of a puzzle and we are trying to put them together so that we can, in the end, maybe find that... if the entity A, the first step, was a drug and then there were five steps in-between but in the end was coronavirus then maybe we can find this connection even if it is indirect. And then of course the next step would not be to give this drug to patients, that would be horrible. The next step would be then to use this kind of, the drugs that we find, in a proper drug screen and then proper cell biology experiments and eventually animal studies before you go to patients. So I am not claiming that by this way we can find a drug that we can give to patients tomorrow, that would be completely overselling what text mining can do. But our goal is to basically find the target that then the biologists can study. So say, hey look here, we found this connection, this is something you should check in the lab whether this is true. At the moment, many research for example a drug screen is maybe just done with a large collection of drugs. Maybe they are screening with three hundred thousand, two hundred thousand compound and just testing them all, but they have no way of saying from before, okay... these ones have a high chance of success, but maybe this is what the text mining can deliver and say, okay... instead of screening these three hundred thousand, look at these two thousand and then maybe among those two thousand is a higher chance of success. And in the same way, say someone has done a big, big drug screen and then they have fifty compounds in the end, fifty drugs, that are potentially useful, so which ones should they go forward with. You know, is it drug one or is it drug two, three, four, five, fifty? There is no way they can test all of them properly in biological experiments. It would take too long time and it would also be to expensive. So they have

a list but they need to prioritize it somehow. But then the text mining again can come in and say, okay... let us look for what information in the text we can automatically find about all of these drugs and then maybe you find okay, drug twenty-two and drug forty-six had actually been mentioned quite a lot in connection with other coronaviruses. Okay, then maybe you should prioritize them for further studies. So this can give some kind of guidance. So this would be an example from drug screening but you can also exchange looking for drugs for example looking for specific information about genes and coronavirus and then you can see, okay... we can find this connection to this and that gene, pops up again and again in connection with coronaviruses, maybe this is something that should be verified in cell experiments to see whether... maybe if you have a specific variant in this gene, maybe you have a higher risk for getting more sick because it promotes the uptake of the virus or something like that. So you can also use the text mining with the medical journals and this is what we are planning to do, for example with the medical journals. There we are going to try... in this free text, look for all kind of symptoms for example so that you can get a proper overview... we have a list of symptoms of course we know for coronavirus... but especially for rarer symptoms we do not have a good overview yet of what patients may present with. So by looking through all of these journals, maybe we can find which symptoms patients come into the emergency room with but also maybe which symptom combinations occur. So it may not be that... a person who comes in with just a cough may have a fifty percent chance of being someone with coronavirus. But a person who comes in with a cough and loss of sense of smell may have a higher chance, and we can maybe find even more complex patterns like that, that can help us identify which patients are first of all positive for coronavirus before we can administer tests. Because if someone walks into your emergency room, you cannot wait for the test results, you need to decide on the spot if this is a potential coronavirus patient or not. And the same for primary care maybe. But also, maybe we can eventually say okay, here are the patients, they have these symptoms combinations and those died and those had a very mild disease and maybe we can do some kind of risk prediction when we have individual patients eventually and say okay... this patient has this combination of symptoms, this means this patient is a high risk of getting sick so he should be monitored much more closely for example or we should maybe give all kind of treatment we have as early as possible because it is a very high risk this patient will get very sick. So this is the type of information where we can extract from with the text mining, which a human person could not see. So any kind of thing that is a more complex connection but also just processing. I mean, there are thousands of medical journals to read through in Sweden alone, so who has time to systematically now sit and read through them and try and find these nuggets of information. I mean, the doctors are preoccupied right now with actually treating the patients, so we need to deliver something to them that can give this information so it is not lost. It does not help us if we find out five years from now what the risk combinations were, we need the information right now and humans are not fast enough to do that.

I: Is this tool up and running right now?

SA: Part of it are but not... so at the moment we have the parts ready that for example can scan for drugs and also some that can scan for proteins in texts, so these parts are running for English text. What we still need to do is work on the… to connect the parts so we only have preliminary versions up and running and then we also want to put this all together and also validate it more thoroughly. Because, as any kind of research method, these methods have limitations and it is important... it does not mean that you cannot use them but it is important to know what errors the system makes

so that you can evaluate the results afterwards. So this we are still working on, so no... the work is not done and with the Swedish texts we are only just starting, so that is not up and running yet. But hopefully it will be soon.

I: And I imagine the main target for this, they are researchers and they are doctors. Do you see any other target groups?

SA: Well, industry of course is also important target group. The information we gather we will release freely so even industry researchers... say people working at vaccine development companies or drug development companies are important target group. But on top of that, I think the same type of tool would also be important maybe for journalists who quickly want to gather information but they cannot read through three hundred thousand articles before writing a story, but maybe these tools can help in that way. Then maybe we can also have public health authorities for example that may have benefit... because they constantly need to review the new evidence and I mean, they also do not have time. So I mean, they need to kind of reevaluate every week whether... did any new measures come out, did some measures prove to not be useful, have we identified new risk groups for example when we want to decide... I mean, risk groups should get the vaccines first, but who is actually the risk group? So this kind of information maybe we can extract. So there are other people so to say, parts of the wheel, that need this information. And on top of that, we should not forget that not just medical researchers are studying coronaviruses and this pandemic. Even people in social science are and people in political science are and people even in law, and they also need to maybe process large amounts of texts. I know projects for example that are looking at connections between social media and coronavirus and these types of texts can also be processed, so we also have some kind of discussions with these type of researchers in Lund and see if we can help them.

I: What makes this software really innovative in comparison with what has been done earlier?

SA: I think it has actually not been done that much text mining for medicine in Swedish yet. I mean, there has been some good work from other groups before but many of these did not yet involve the latest technological developments that have been made in text mining. So the latest type of AI-models for example, they are currently not available for Swedish medical language at all. So this will be a big innovation, if we can get these tools running many other people in Sweden could use it for their research afterwards, even those types of research that are not related to coronaviruses. So I think this could be a kind of a general boost for text mining in Swedish and even when it comes to English... there is of course much more published in that area but many of these systems, they are not very mature yet and I mean by that, that maybe specialists can use them but the target groups can actually not. So our goal is to develop tools in the end that would not need a collaborating entity that is specialized in text mining. So that we want to make a tool that a virologist can use even if they do not know a specialist in text mining. So this is kind of our goal and this is something that is lacking at the moment. Many of the tools we have are only accessible to really specialist teams.

I: So how new is this kind of text mining technology?

SA: The big breakthroughs have really just come over the last couple of years. So a big part of it was that we established something that is called transfer learning where what you do is that you would first train an artificial intelligence model on a very large collection of text. And that could be, for example, the entire Wikipedia and a thousand books. And that kind of gives the model a general

understanding of the structure of language or the structure of Swedish medical language, let us say, if it is trained on a Swedish medical text collection. And then you can kind of, from that, adapt that to specialist tasks. So for example, finding specific types of symptoms in the texts or finding relationships between drug and disease, so this kind of a specialist task. But it helps if the model basically... if you can make use of a model that already understand medical language, so then the training goes much faster and also you can train it with much less data, which is important because in order to train a model for, say recognizing symptoms, you need to have a collection of texts where researchers have highlighted, here is a symptom, here is a symptom, here is a symptom. And making this of course takes a long time. So if you have a model that has a general understanding already, then it is fast...

[audio cuts out]

I: So that was the innovative or the technology... the history of this text mining technology… what you were describing, is that what is known as deep learning?

SA: Yes. So the type of text mining that we are doing is deep learning-based text mining and this is a rather recent innovation for text mining. Before that a lot of work was done doing text mining based on matching specific keywords, just with long lists, just trying kind of to match. For example, you have a long list of all the symptoms that you can think of and then you try and match that. Or you would have some kind of rules, all the words ending with a V are antiviral drugs for example, these type of things. But of course the systems were much more prone to error because for example... just to give one example, there is the word lamp as the one that makes light, but there is also a protein called lamp. So if you just match lists of texts then the tools could not distinguish and they would, every time someone describes an experiment with an UV-lamp, they would say... this is about this protein, which is of course wrong. Whereas the deep learning-based models that we have today, they can take the context into account of a sentence. So they would for example be able to understand the difference between, I will deposit my money in the bank or I will sit down on this bank... on this river bank I mean... so this is something that deep learning-based models can do. So they can distinguish and it is the same... they would be able to distinguish, this is a protein from this is talking about the UV-lamp. So this is one of the recent advancements for text mining. And also the use of pre-trained models that are... for example, we have models thar are already trained or used in scientific English language. So someone has taken a huge collection of scientific English texts and trained a model in general language understanding. And this we can now use as a starting point to just give it a specific task and it’s something like this we also need now for Swedish. So that does not exist. We have the first Swedish language models, have recently been developed, but they are for general Swedish language, they are not for medical Swedish language. So this is something that we also want to help develop so that we can make use of the latest technologies even for Swedish, because with the smaller languages we are always lagging behind. There is so much research done for English but there is only so many people that do research for Swedish of course, and we just want to help catch up with the latest developments in English so that we can make use of them for Swedish.

I: What has been... we were talking about deep learning and what is the technological background in this? What has really made this possible?

SA: I think the reason why we have seen this explosion in deep learning, there are a couple. So first of all in the past years, we have produced an enormous amount of texts and also other type of data that is just now available. If we would compare what has been generated in biomedical research data in the last ten years to… it would probably exceed what has been produced in the hundred years before. So it is just an explosion of data that we have seen and deep learning requires often a large amount of data to train the models, so this helps a lot. The other thing is that there have just been some conceptual breakthroughs in how to train these models better and how to build these models better, that have come over the last years. For example this now prevalent use of transfer learning where we make use of pre-trained models that were developed for another context but can then be quickly readapted, this is also a rather new development. And the other thing is that, of course the computers have gotten better. So a project just like ours, it would not have been possible to do for a research like ours fifteen years ago. I mean, the hardware would not have been there. I mean, maybe you could have done it but it would have taken so long time it would not have been meaningful. So for example, many models in the type of artificial intelligence-models, they are very large when you kind of store them and so on and do all the calculations on them. They can have millions of small calculations that are made inside these models. So they did not fit into the memory of the old computers for example. So then people used types of tricks and tried to split them up in different pieces but none of this worked as well, so we have just gotten better hardware as well. I think these three things together have kind of made this huge explosion in deep learning which is not just for language technology but also for image analysis for example, which also use the same concept of this transfer learning that you use a model pre-trained on one domain but then quickly adapt it to a new task. And this insight has only come a couple of years ago.

I: And in this project, are you using hardware resources from the SciLifelab?

SA: At the moment we are primarily using the Swedish National Computing Infrastructure, SNIC. So we have a network of supercomputers in Sweden that we as academic researchers have access to for meaningful research, so not for our private fun but you apply for access to these, describing the project that you want to do. And then you can use these resources, so we at the moment are doing most of our computing both in the supercomputing center in Lund, which is called LUNARC but also in Umeå which is called Kebnekaise.

I: And your project organization, who are working in this project and what kind of disciplines are represented there?

SA: So, like I already said, I am a cell biologist. I am probably the only person with an extensive medical... biomedical training in the team. So the other people we work with are people from computer science. So they are specialists in machine learning from the department of computer science in Lund for example. We also work with specialists in language processing that are coming from infrastructure in Lund that is called the humanities lab, so it is kind of a core facility that does computational language technology. We also have some kind of people with an engineering background in my group, because in the engineering fields people program a lot so we have people that maybe do not come from computer science but come from engineering and have extensive computer science skills, I would say. Then we also work with... actually, I am not the only medical person... we also work with clinicians of course. So people that actually work in the hospital. So for the Swedish symptom text mining, we work with a doctor who works in an emergency room and also

some of her colleagues who are specialists in public health. So we kind of try to... we have also had some input from actually people working with veterinary public health, the people work at statens veterinärmedicinska institut... I am not sure what the correct name is. And also we have gotten input from people who work with systematic scientific review. So there is an organization called Cochrane, basically they try to summarize the state of knowledge currently manually. So we have gotten some input from them because we are trying to basically make their life easier with the tools that we are developing. And then we have also gotten some input from CEPI which is an organization that coordinates many of the vaccine development that is currently going on. So we try to get input... we also have some mathematicians we sometimes talk to. So it is very diverse team and I think it is absolutely necessary because otherwise, like I said... none of us can grasp all the aspects of the project, it is just too complex. So it is important to get input from someone who is a specialist in epidemiology and from someone who actually writes medical journals for example, and so on. And for the genes and the drugs and stuff, it helps of course a lot that I have experience in making a, kind of, scientific experiments in the biomedical domain. Because then I can say okay, this is not a protein or this is not a chemical this sentence, it sounds like one but it is it not and these kind of things. And of course, you need the people who actually program extensively and who can train these models efficiently.

I: And you told me before you are also cooperating quite a bit with the other projects in the SciLife-program. What are those projects?

SA: One of the projects we especially collaborate with is the project of Darcy Wagner and they are specialists in lung biology and lung tissue engineering. So they are also part of this... so basically, they are the ones who do the biological experiments to then verify drugs and develop new model systems to test our treatments and so on and to understand the disease better. So they generate of course a lot of experimental data that we are trying to help analyze and we are also trying to deliver information to them that will help guide their experiments, so to say. But I am sure that over the course of this national program, we will have more collaborations. There is for example another group that develops a platform which makes trained artificial intelligence-models available for the general user. So we also have already discussed with them that once our systems are running, we will kind of work with them so that they can put them on their platform to make it easily accessible. Then, as I mentioned, we also helped develop these data portals so whenever we find input and resources or datasets, then we always deliver them to the data portal and we have also been extensively kind of... been involved in discussing how the layout of the website should be for the resource section and so on. So this is contributions we are trying to make and we will also have some additional discussions now with the virologists to see if we can help them as well. But it is of course also a question of manpower. I would like to help everyone but there are not enough hours in the day.

I: So what would you say have been the biggest challenges?

SA: I think… The biggest challenge is probably... for the Swedish part is the getting access to the data. Because that just simply, you know... it has to go through the appropriate way and that simply takes time. So even though you are eager to start and ready to start, you just have to wait for the ethical approval and all of this stuff. And I know this is also a big challenge for other text mining programs that work with medical, clinical data. Or really any other clinical data, so even people who do not

work with text mining. So data access is a big challenge. Another challenge is that actually... within this supercomputers that we have, there is currently no setup for GPU computing with sensitive data. So with the type of data we are using... so we have to kind of build or own dedicated infrastructure or hopefully maybe we can also borrow some infrastructure from people here in Lund. But my group was not set up because we worked with literature before and that was not sensitive data so we were not yet set up for sensitive data. But we are getting lots of help from the people who are more used to this so that is... I am confident something we can solve. But this was also a major hinderance and I think then, from the... but I think this is maybe for all the people that now sit and work at home, it is a transition to move to a completely digital work life. So because the guidance has been that people who can work at home should work at home, we have actually taken this very seriously and my group has been working at home since the spring. This means that I have supervised students in these projects that I have never met in real life. So you know, we make it work but it has taken some time. We already had digital tools in our group of course and the staff I have, they are good with computers of course, so that kind of makes it easier than for many others. But often you get ideas while sitting and chitchatting at coffee for example, you get new ideas. And this is missing so we are trying to recap that, we now have a Zoom-lunch together, like a social Zoom-lunch and trying to get a little bit of that back. But it is of course not... we do not do that every day and it is not the same of course. So this has also been kind of... a little hindrance. But at the same time, we have... through the digital work life... now gotten connections you know, to people that we were not connected with before. So it has not just been hindrance, the opposite I would say, we have benefited a lot, we have learned a lot from others. We have gotten a lot better at networking, we have gotten a lot better at exchanging tools, so overall I think science will get also a huge boost. I think science in general will get a huge boost from this pandemic, not just because a lot of money has been spent but because people have found new ways of interacting and new ways of collaborating and there is more openness maybe now than there was before. And I really hope that we can retain that even when this pandemic maybe eventually dies down and make use of this new kind of connectedness and new way of working together more openly. Even for other kind of issues, and I am confident that this will happen. I also think that we have gotten a lot better at, for example, having digital meetings and teaching digitally and that can also make it easier for people who maybe for other reasons cannot access physical meetings. It may be that you do not have time to go to a seminar in Lund because you are working in Malmö, it can be as easy as that. And there is a super interesting seminar but you cannot afford to get away for two and a half hours but you can log on online and maybe we have students who are located in other countries who can now take part in educational activities even though they may not have the possibility to be physically in Sweden. So I think we will have lots of benefit from what has happened this year in the long run as well. And I guess that is so with many crisis that many crisis can really be a motor for change and a motor for innovation. These are all things that have been in place before to some extent but now they have actually been made use of properly. So maybe... it might also be good for the climate because maybe now people who travelled a lot to meetings see that maybe a part of those meetings can be replaced by digital meetings. And in the end that will maybe reduce our carbon footprints. I mean scientists are among the groups of people that travel the most, that is by nature because we meet up at conferences and we have international collaborators and some of this travelling will still be necessary even in the future and we can see that it hinders us you know. We cannot make new connections as easily if we do not meet up physically at a conference for example. But I think overall maybe we... if we just want to talk

to one of our long term collaborators in Stockholm, maybe we will not actually bother to travel for every meeting. So I think lots of good things can come out of this if we try to make use of it.

I: I think you are absolutely right. You talk about the effect in the long run. What do you see, the effect in the long run of text mining?

SA: I think we just will be able to make use of the information we have so much better. At the moment for example, there are thirty million scientific articles published in the medical domain. Sometimes information relevant to a cancer researcher may be found in an article from cardiology. Right now it is very unlikely that this information will cross the domain. I think text mining can help put all of these information pieces together and I hope that, as a result, we will actually get a more comprehensive insight into what is actually going on in diseases or in cells for example. So not just get a snapshot but kind of put the information better together in a better way. But that we also just can do things faster. So the same things we could do today in months, maybe then we can do in a day. And it is the same way that... when my dad was a student, he used to go to the library when he wanted a scientific article. So he would actually have to physical go there, then find that article in the magazine and then take a copy and then walk back. So he would maybe access, I do not know, a couple of articles per week. Whereas I now, in my PhD accessed hundreds of articles in a week sometimes. Kind of just scrolling at least through the summaries just because I have the internet, and that made all the difference. And I think these type of tools, it is the same way as what the internet has done for us. I think artificial intelligence and difference aspects of it, including text mining, will make our life easier in many ways the same way as the internet has done. So it will not be something that is used by specialists, this is something that will be everywhere from customer services to clinical studies to cell biologists to people in politics. I think everyone will make use of text mining. And actually, in a way, everyone is already making use of text mining because people are doing Google search and that is also natural language processing.

I: That is true.

SA: And they are talking to their Siris and Alexas and that is also natural language processing. But all of these tools have gotten a lot better over the past years and they will get even better over the next five.

I: Before we wrap this interview up, is there anything you would like to add that we have not been talking about?

SA: I think one thing is that... what anyone working with artificial intelligence and especially if they are working also with medical data should always think about is the ethical aspects. So when we work with text mining, we can automatically extract a lot of information, potentially also sensitive information. And I just want to... I think it is important that in this disciplinary work we do, it is not just about bringing medical people and computer science people together but it is also about bringing computer scientists, medical people and people who are interested in ethics and social science together. And a lot of this is what we are trying to do in AI Lund. So we have this network in Lund where is completely interdisciplinary, we are covering all the faculties and we have a lot of artificial intelligence themed events together and we talk a lot more to each other than we used to do and I think this will also be a huge change for these fields, that these aspects... so the humanities and the social scientists are also included and I think this is also something that we have realized that

chatbots can get racist if you are not careful what you are training them with for example and these type of things. And I think this is important, that we always keep an eye out and also that we need to assess the quality of the data that we train the models with. So are they biased these texts that we are using, are they just untrue... because we also have some faked scientific articles for example... so can we also use these tools to distinguish good data from bad data for example and only use the good data to train our models and these kind of things. I think this will be kind of things we are seeing more in the future and I think this should also involve the general public and patient groups. So if we use text mining prevalently everywhere, I think people should have some kind of basic understanding so they can say yes or no, as ever they want, to being involved in such a study. And that is, I think, important that we as scientists come out and try and explain but also... it not just us who come to the general public, it is also the general public who should come to us and ask the questions. So ask the easy questions, ask the hard questions, there are not any stupid questions for sure. So scientists are just ordinary people, so if you are a person who reads about a scientific discovery and it blows your mind and you find it so fascinating and you want to know more about it, sometimes just drop a line to a scientist you know. Most scientist I know they are happy to answer, so they are happy you know... invite them to your school if you think, this is something that my school class should learn about. I think, we should work more closely together, kind of the scientist and the general public but this is not just the scientist coming to the general public, it is also the other way around. I sometimes get e-mails, I recently got an email from a high school student who wanted to do a project and I gave some guidance and this was just fun you know. Whenever we have time we will try to be there.

I: We are trying to bridge that gap.

SA: Yeah, right now I guess.

I: Exactly. Sonja, thank you very much for this interview. It has been a pleasure.

SA: Thank you so much for giving me the opportunity. You know, having a chance to talk about work for people who are not scientists is lovely, so thank you very much.

END OF TRANSCRIPT/LW

Redigerad av Peter Du Rietz