Verifying Software Assurance with IBM's Watson

featuring Mark Sherman as Interviewed by Will Hayes

Will Hayes: Welcome to the SEI Podcast Series, a production of the Carnegie Mellon University's Software Engineering Institute. The SEI is a federally funded research and development center sponsored by the U.S. Department of Defense and operated here at Carnegie Mellon University.

My name is Will Hayes, and I am a principal engineer here at the Software Engineering Institute. It is my pleasure today to welcome my colleague, Mark Sherman, from the SEI's CERT Division. Mark serves as the technical director of CERT's Foundations of Cybersecurity work.

Welcome, Mark.

Mark Sherman: Hi, thank you for having me here.

Will: Before we get started, could you tell us a little bit about your background and how you found your way to us?

Mark: Sure. I started actually as a research faculty here at CMU, and we formed a spinoff. That spinoff was later acquired, and I left that to come back to CMU. Here I am at CERT, at the SEI. It is a really exciting place to be.

Will: Could you tell us a little bit about the focus of your work? I know you have got some really neat stuff going on.

Mark: One of the primary research areas of the group that I have is how to build software securely.

That has a lot of elements to it: all the way from understanding threats, to understanding the risks those threats have, how you model them, how those then generate requirements, how you then have an architecture to support it, the design to support that, the actual implementation, and whether that code is working well, and the testing and evaluation and deployment and updating of that software—the entire lifecycle.

DoD has something very similar to that. That is what they usually phrase the acquisition lifecycle, which carries out the same kinds of activities although frequently labeled in different ways: Milestone A, Milestone B, material analysis, and so on. Our group is looking at how to make that better so that we create better software that is more resilient to attack as opposed to afterwards having to deal with all the attacks.

Will: While you have a very deep focus in technical matters relating to software, the application of these important concepts is very broad it sounds like.

Mark: Well, it is for people who are trying to develop software, again, resilient to attack; whether it is for software that goes into an airplane, software that goes to electric meter or CRM systems that conventional companies would run.

Will: It is everywhere.

Mark: That is pretty much like software.

Will Hayes: The very interesting twist we have to talk about today is you have got a project that involves IBM's Watson in the space. Could you tell us about Watson a bit?

Mark: Sure. Just to take a little step back about where Watson fits in and why it is interesting. To do this evaluation of how well you have understood your requirements, your threats, and your architecture, there are usually a lot of documents that are produced along the way. Especially within the DoD environment, there are a lot of very structured pieces of information that programs are required to assemble to provide evidence that the system will do what it is supposed to do and is resilient to attack, its performance, and a variety of other kinds of attributes.

Answering questions from a set of documents is quite daunting. A friend of mine at MITRE—I can't tell this first hand, because I did not do this—but the friend of mine commented that the specifications of what goes into those documents takes a week to read. The documents that are produced are just enormous. Many times to carry out this exercise what happens is that DoD will bring together in a room the experts in each of the individual areas the documents talk about. When they have questions they will sort of put their heads together rather than actually going back to all these documents that have been produced.

The thought was that Watson could read all of these documents on behalf of a program manager or some other interested party. Then you can start asking questions of Watson, *Is this a true thing about the system that I am looking at?* because we spent all this time and effort building this information.

Will: So the premise going into your project is the rich and well-established structures that exist for specifying this information would serve you well?

Mark: That is what we were expecting—that we would be able to use Watson to exploit that and provide assistance to program managers.

Will: Just as we watched Watson compete on Jeopardy, what would it mean for Watson to win in this environment. Tell us about what good performance means in this application.

Mark: Coming up with a precise metric was challenging. Much of what we were trying to do was not so much build a product tool that an actual program manager could use. We are a research organization focused on understanding technology better. The underlying question that we were trying to answer for the DoD—I think we did—was *Is Watson technology usable for this kind of application that has a variety of aspects to it*.

First of all, can people who are—I use the *phrase journeymen programmers* as opposed to *Ph.D. researchers in artificial intelligence or cognitive processing—could journeymen programmers build a Watson application?* After they built the application, was it something that actually provided some value? Did it actually help? How much of what you were trying to accomplish could it actually help?

We carried out that exercise, and what we learned was a couple of things. First, that yes, people who—in this case we had undergraduate student programmers at CMU, so that sets the context—took approximately one full-time year of effort to build the system that we did, obviously spread out over time. Actually we had multiple people working on it, but it is approximately in total one FTE [full-time equivalent]. Understanding Watson was a straightforward exercise. These were not particularly students specializing in AI [artificial intelligence]. They were computer science students.

Will: Nor defense acquisition…

Mark: They did not know anything about defense acquisitions. We did have subject matter experts as part of the team. They just came from the SEI, from my team, because obviously we understood this. We helped identify what was the important information in documents, but once that was done by an expert, then the students were able to generate all the necessary files that were required by Watson.

Will: While we observed a moment ago that the documents are written with a very well-established and detailed structure, you had subject matter experts that helped to codify the information?

Mark: One of the things that we were looking for was, what were the perhaps unknown issues in building a Watson application that you might have not picked up if you just watched Jeopardy and said, *I want one of those*.

One of the things that we learned was that Watson works best on small pieces of information. Just to give you an example, there is something called the CIA [World] Fact Book that lists all countries and populations and capitals and so on.

You could use that and teach Watson that the capital of France is Paris, the capital of England is London, and then go ask it, *What is the capital of Germany?* Watson will come up with Berlin. But if all you did was feed it the book, all it would come back to you with is the book. So if you said,
Show me the capital of Spain, it would give you the entire book back because in the book it has the capital of Spain.

Will Hayes: And it wouldn't know about Brexit or recent challenges in the economy there.

Mark: Only whatever you feed it. To make it valuable—to use that example again—you actually have to fragment that document into pieces. For example, you need to take out the sentence, *The capital of Spain is Madrid. The capital of France is Paris*, and have that be an individual document that then Watson can go find.

That pre-processing is what the subject matter experts are really for. It is to say, *Given this document—so an acquisition document is 100 pages long—how do you want to structure it? What pieces do you want to take into different elements, so that when you ask a question it can return a precise answer, rather than just giving you the whole acquisition document back*. You know, giving you back the entire CIA Fact Book. That is not what you want. You want to know the capital of Spain.

Will: We have heard amazing things on television about the way Watson is changing the medical field by ingesting huge amounts of data in recent publications about medical journals. I imagine there is a pre-processing step of this sort that is supplied there?

Mark: Oh, absolutely. One of the other elements we learned about Watson is that there is a community of developers. IBM helps them as well, so it is not just application developers. And they share a lot of information amongst themselves about how to build these applications, which we tapped into, as is fairly easy to do by frankly anyone who wants to do this.

The construction of these fragments—is the phrase we used in these documents—and later the training against those fragments is the overwhelming amount of effort that goes into this. I gave you an approximation that we spent about a year of person effort of building the system.

I would estimate that is about three weeks or so of the subject matter experts putting together the guidelines, maybe two weeks or so or three weeks at most, of actually building the application, and all the rest was data preparation to go into Watson.

Will: Interesting. Is there automation on that front end that is developing on, *How do we do that pre-processing?* Is there progress there?

Mark: That gets to another I think important lesson that we learned. You mentioned the medical application. I am not as familiar with the structure of the medical application, but we did talk with other people doing other applications. A recurring theme that we validated is that Watson provides one kind of cognitive processing on the information. There are lots of kinds of cognitive processing and other kinds of artificial intelligence that can be applied to natural language processing. Most of the commercial systems that we worked with—again, I did not deal with the particular medical systems, but it also includes the Jeopardy system—include multiple additional kinds of processing, in addition to what Watson is.

I think now is probably a good time as well for those who may not be aware, Watson these days is actually a whole collection of technologies and products that IBM offers. When I use the word *Watson*, and what we focused on, and what was used in Jeopardy—again, the thing that caught everyone's attention—is the question-and-answering system. There is a whole variety of other kinds of processing that is available under the Watson brand these days. We didn't use those. We just used in particular the highly publicized question-and-answer subsystem of Watson.

Will: The privilege of being at a university, such as the one we are working for, gives us opportunities to interact with leaders like this. What is your vision for where this can go? If everything turned out exactly the way you would like it to, what happens for the world?

Mark: Well, what we were again focused on was, *Can this be used?* The short answer is, *Yes, it can* with the caveats that you need some subject matter experts to help to take apart the documents into pieces that are meaningful that give meaningful answers, that you give training questions that are relevant to the people who want to know information. In the case of Jeopardy, they taught it about puns. Well that gives it information about puns, but not necessarily information about acquisitions.

Will: Or cybersecurity.

Mark: That is another important element. The fact that you should expect—if you want a high-quality production system—to extend the cognitive processing. The way that we, for example, did the fragmentation and generated the questions for the training of the system was I will say *algorithmic*. To use the CIA Fact Book example, it is pretty easy because that book is structured there. The title page gives the name of the country. Then it has "capital:," and it has the name. So it is very easy to basically use a search or grep commands that just find those particular pieces and extract them out. In a more realistic example, you probably would have to apply additional natural language processing to extract the information out of that.

Again, many of the partners that we talk to who are building Watson applications did have that additional pre-processing to create the documents that would go into Watson. In some cases, they had additional cognitive processing after Watson. So Watson would come up with—and this is typical—Watson comes up with a list of answers. The 10 answers, and then how do you rank the answers? Watson will come up with its own ranking, but then you have to decide whether you like that ranking and how you might rearrange that ranking, and post processing might lead you into other things.

Will: You have established an initial test bed, and had success with it. You have invested a lot in training these components of Watson. Is this something that persists and can be used over time, or does it require refreshing and retraining at a level that we have not supported?

Mark: Again, our experiment was to answer a specific question about technology maturity. In general, for this to be useful, you would like Watson to continually ingest new information—whether it is about the acquisition program or baseball scores or medical research, or whatever it is that you are looking for—which is why the generation of how you pick apart the documents and the training questions itself needs to be automated.

You just can't have someone decide they are going to spend a weekend and write down the 10 questions that they want the system to be able to answer as samples and go on. Again, this seems to be a recurring theme. We used automated processes for that as well. We generated, I am trying to remember this, tens of thousands of questions.

Again it is not someone, you and I, saying, *Gee, I wonder what would be 10 or 20 interesting questions they might ask.* You need to give it a huge amount of these kinds of things. All that has to be automated. That again is natural language processing. We actually took advantage of the fact that we are here at CMU to help drive that testing and that thinking and the use of people to help train the system.

Eric Nyberg, who is a professor here on campus actually was one. He and his students were among the people who helped IBM build the original Watson application. We were lucky to have

him involved with our project as well. He sat down and used the system and gave us advice on, *This was not trained well, and this is why you can do it better. You have to sharpen up this particular element*. That was actually a very, very good circumstance for us.

Will: You have got a recent publication and a tech note that covers this area?

Mark: Yes.

Will: Can you talk about that for a moment?

Mark: We have both the video from the SEI Research Review, which gives a very I think it is 5- or 10-minute brief overview. It gives you the highlights of what we did and the conclusions that we learned. It has some measurements as well. We can talk about those later, if you would like.

Then we also have a more complete technical note. We also have basically what I call the half an hour presentation that we have given at some conferences. When asked by customers we do it as well. We have had some parts of the Department of Defense who have been interested in using this kind of technology for their own needs.

Will: Those are things people could access from the SEI's website by getting to your digital library?

Mark: My understanding is that we publish all of our research review videos.

Will: A search for your name usually gets a complete listing of what you published and access that way.

Mark: The tech note will be there as well. The reason why you see me frowning a little bit is that one of the talks was given at AAAI FS, a conference in D.C. I am not quite sure where they published their materials, but we certainly can make it available to anyone who wants.

Will: Yes, people can reach out and talk with us. So what is next, what is your future here?

Mark: Again, this was intended to be a very scoped project. We wanted to make sure that what we learned was of some kind of recurring value. The reports that we gave really give guidance to programs who say, *We want to use this technology to solve a problem. What can the technology really do? What should we expect? What kind of people do we need to put on the project? How much time is it going to take for them to do? What are parts going to do? What parts have to be done by some other kind of technology?*

The specific work that we did, we actually licensed to another company, SparkCognition, down in Austin, Texas. They built a Watson-based application called Spark Secure, which uses Watson for cybersecurity.

They licensed the technology that they could including their product to add additional corpus of information, which is really the hard part of the system into the corpus that they themselves already developed containing cybersecurity information.

Will: So with your initial experiment done, you can see this work progress in other venues.

Mark: The specific details of what we did moved into the commercial world, which is again one of the things we try to do as an FFRDDC. Then the knowledge that we gain provides capabilities to us so that if the DoD needs help in these kinds of areas, we can give advice and assistance in that particular area.

Will: Understanding the data coming in is probably a source of keen insight for us that has application in other places as well.

Mark: Maybe in the Watson technology, but we started as subject matter experts in building secure software and acquisition practices. So Watson was not really a learning exercise in that area, but it was arranged for us to use our existing expertise in a new kind of technology.

Will: You mentioned some measures of performance for Watson. Could you give us a little bit of detail to kind of give people a teaser for the tech note?

Mark: Sure. In the report that we generated, we provide data on how many, how much data we took in, how many training questions were necessary, how many different concepts were necessary. We also did some studies of precision and recall.

There are two measures of how well the system finds the right stuff and how much it does not find the wrong stuff, and measure that against other systems as well. Just to give people an idea of the performance of the application.

Will Hayes: So you are publishing empirical data on the performance of the system, not just talking about the concepts here. People can access that for the future?

Mark: Yes.

Will Hayes: Mark, thank you for joining us. This has been a very interesting conversation.

Mark: Thank you for inviting me. I enjoyed it.

Will Hayes: As always, this podcast will be available for download on the SEI's website, sei.cmu.edu/podcasts. If you wish to access the work discussed here, including the presentation that Mark mentioned, you can search for his work on the SEI's website by simply typing his last name, S-H-E-R-M-A-N.

This podcast and its transcript will be available also through Carnegie Mellon University's iTunes U site as well as the Software Engineering Institute's YouTube channel. Thank you very much for joining us today. If you have any questions, please do not hesitate to email us at info@sei.cmu.edu. Thank you.