DAISY THE GREAT: MUSIC

CRAIG: Hi, this is Craig Smith with a new podcast about artificial intelligence. This week, I talk to Ken Church, a pioneer in Natural Language Processing whose use of statistical models on part of speech tagging and more revolutionized the field. We talked about his early days at MIT, about explainable AI and how the Holy See played a role in his probabilistic approach to NLP. I hope you find the conversation as enlightening as I did.

CRAIG: 00:45 You are regarded as one of the pioneers of natural language processing. And I'm interested in hearing about people's journeys, how they became the people they are - not only the research.

KEN: 01:00 Sure. So I was an undergraduate, and a graduate at MIT, and I got into AI in the mid-seventies. I took one course from Patrick Winston and I was hooked. He is still teaching there.

CRAIG: 01:15 This is about what year?

KEN: 01:16 ‘74

CRAIG: hmm.

KEN: 01:18 Right. And that that was undergraduate and I graduated in ‘78 but I stayed at MIT and then I was in an expert systems group, medical decision-making expert systems group. And in those days they were worried about - topic has come back again - explainable AI. So, the concern was that the doctors wouldn't accept the answers from the machine even if they were right, if they couldn't be explained. So we had expert systems that were coming up with opinions, but they then needed to explain the opinion to the doctor. And what counted for an explanation in those days was a trace of the program, like, you know, I did this because I did this because I did this because I did this.

KEN: 01:56 And it started to be like, you know, that whole, the rat, the mouse, the cheese, ate, you'll have to go look it up, I don't have it quite right. But the way the computer thinks is with lots of recursion and the stack trace is just not an easy explanation for a person. And Chomsky was at MIT in those days. And so I started taking some of his classes and he had, uh, there's this thing called the Chomsky hierarchy. The simplest case is finite state where you use a finite amount of memory and the arguments that work best with people sort of follow that. But the computers were using context free, which requires a big stack, a lot of memory to keep track of it all. And this is actually very hard for a person to follow, that it's just not an intuitively sensible argument.

KEN: 02:45 So my Master's thesis was about, was trying to sort of rewrite the arguments that the computer used into an organization that would work better with people. Chomsky's argument was that the simple methods that, that are now popular again in machine learning, were fundamentally inadequate because they couldn't capture the interesting parts of language. But using his same theorems, you could sort of say that in fact, maybe it's that if we want to make a computer explainable to a person, we need to make the arguments simpler and easier to understand. So that's how I got into it. But I never got back to the explainable AI part. And nowadays, in fact, I recently wrote a paper called, 'I did it, I did it, I did it, but.' And the idea is we get better and better numbers all the time. Results are better and better and better.

KEN: 03:45 There's more and more papers. There's all this excitement. Results are really getting good. But nobody knows how they work. Even the people who write the programs. And so this idea is, it used to be, 'I did, I did, I did it, but I'll be damned if I tell you how, it's trade secret.' Okay, now it's, 'I did it. I did it, I did it, but I don't know how.' And so we've gotten more and more publications, but less and less insight. And it seems like insight is no longer required or even expected or even appreciated, that all we want to see is better numbers. And I think that there's a push back on this that sort of says that it isn't enough to get the right answers, but you have to be able to back it up with an explanation.

CRAIG: 04:15 This is particularly the problem with deep learning.

KEN: 04:18 Especially there. And then you get into problems like with the regulators. So there's a recent bestseller book called 'Weapons of Math Destruction.' It's got one argument after another after another about this problem of explainable AI. So this whole question of we're getting the right answers, or in that case, she's even worried there's some snake oil salesman that are selling boxes that don't even work. Then governments have to make lots of decisions and other people like who gets into a good school, who gets a loan, who goes to jail. You know, a lot of these are thankless chores and they would love to delegate it to an automaton and then the snake oil salesman says, 'I've got a proprietary thing. I won't tell you how it works,' and they don't even claim that it's safe and effective. I mean, so the FDA has rules about how to decide if a drug is safe and effective, but here we don't.

KEN: 05:10 So she's really concerned about this and what are we going to do about it. But, my criticism of her book is that I agree that there's a lot of concerns here, but I'd like to see a constructive suggestion forward. What I like about, you know, the FDA thing is that we agree on how to decide if a drug is safe and effective, but the FDA process doesn't actually require you to understand mechanism. You don't need to know why it works. There's also, I think some real questions about what the regulator is supposed to do and maybe we need to rethink some of this. So, redlining is illegal. You're not allowed to decide who gets a loan based on Zip code, but are you allowed to use a variable that's correlated with that? Okay. All right. And if not, is it okay if you don't know. But then the question is if you don't know, how the regulator is supposed to know.

KEN: 06:06 So I think there's some questions society has to work through here. But you know, on the other hand, the stuff is getting the right answer on a lot of things. So I think that it's good to be thinking about these questions. They're important questions, but sometimes you do get the right answer and you don't know how.

CRAIG: 06:17 And sometimes you get the right answer and you don't see the bias or the unfairness in the answer.

KEN: 06:23 Well, all that's in there, so that, I mean, one of the things in her book is that at best all you're going to do is repeat the mistakes of the past. Right? Okay. So you know that that's definitely a concern.

CRAIG: 06:33 Where were you born and brought up and what did your family do?

KEN: 06:38 So, that's all good questions. So my father is a professor at Brown and he moved there - he was a, he went to graduate school at Harvard when Skinner was there and he taught psychology at Brown for many, many years. In fact, he only retired a couple of months ago.

CRAIG: 06:50 Wow.

KEN: 06:51 Yeah. So, he's been, he started in the 50s, I think probably a year or two before I was born, and then retired just now. So he's been there forever, you know. So I'm, I grew up in Providence.

CRAIG: 07:08 In academia.

KEN: 07:09 Yeah. Right. Yeah. And then I went to Classical High School and then I went to MIT and stayed at MIT for graduate school. Then I went to Bell Labs for 20 years.

CRAIG: 07:20 That's right.

KEN: 07:20 Yeah.

MUSIC: 07:21

CRAIG: 07:28 And NLP, the state at which it was, you were instrumental in shifting the, the strategy.

KEN: 07:36 I sort of led the movement towards statistical methods.

CRAIG: 07:40 Yeah. And for people who are not familiar with NLP, what was the method prior? And then ...

KEN: 07:47 So the first statistical paper that I published, the first coming out of my, you know, of the new way of seeing things was a paper that I wrote about part of speech tagging.

KEN: 07:59 Now I think everybody kind of knows what a part of speech is - a noun and a verb and an adjective. This is not a practical problem, you know. But it was a very simple to describe kind of problem. We all have strong intuitions about this. You know, maybe it's part of a solution to something later. It got a lot of citations 'cause I think a lot of people think that it's important for that but. And there were a bunch of things I was concerned about. When I got started in this field, my professors said that it was no longer possible to get a PhD doing anything about parsing and parsing's harder than part of speech tagging. And 10 years later, I, I write this paper on part of speech tagging and I was really kind of nervous about it because the field had declared success on all the things we could do and was working on things we couldn't do.

KEN: 08:50 All right? And here I was working on a problem that was much easier than many of the problems that they had declared success on. So a lot of what you did in this statistical stuff, especially early on, was to address really simple fundamental problems. They're not quite, you know, they're not into applications like reading or translation or anything, but they might be a means towards an end. All right? So, I'm attacking this really simple problem. And then what I showed was that the methods where we had declared success really didn't work, and these very simple ideas did, and I can go through some of those examples. All right. So, here's the argument I was using. I had actually tried to put together, using the old methods, a tutorial for my colleagues at Bell Labs on how to do parsing, which is harder than part of speech tagging.

CRAIG: 09:39 Just explain parsing ...

KEN: 09:40 Parsing is like I think what we called in elementary school 'diagramming a sentence.' So, instead of just saying where the nouns and the verbs and adjectives are, I want to know where the subject, the verb and the object is and maybe how the noun phrase - you've got a sentence, it's made up of a noun phrase and a verb phrase and which words are the noun phrase, which words of the verb phrase and then within the verb phrase, there'd be a verb and an object and an indirect object and a direct object and, how do you do all this sort of case assignment you would do if you were translating to Latin, right?

CRAIG: 10:10 You mean automated.

KEN: 10:11 Yeah, so that you know the program should input a text string and output the parse tree would be a description of where all the phrases are and how they fit into a tree. And that that would be the answer to these questions of things like 'what's the subject, the verb, the object' and 'what's the indirect object' and 'what gets ablative case and what gets accusative case and what gets dative case and genitive' and all that. All right?

KEN: 10:35 Any rate, so, much simpler is just 'what's a noun and what's a verb.' Okay. And so we had declared success on all of this. In fact, we couldn't do any of it. So I was trying to show, I started with like a four-word sentence: I saw a bird. But I happen to have Webster's dictionary online. So I hooked up the program to that instead of what we used to do is to make up a dictionary and I'll say 'I' is a pronoun and 'saw' is verb and then ...

CRAIG: 11:05 You would just label all, label these things.

KEN: 11:06 ... label these things. So, I'd started with a dictionary, just said, here are the words for each of these words, these are the parts of speech they could have and then I could start to say that, uh, you can write a context free grammar, or you could say a sentence goes to noun phrase-verb phrase, and a noun phrase goes to a noun, a verb phrase goes to a verb, followed by maybe some noun phrases and start to write this recursive expression.

KEN: 11:28 Then you hand this to my context free parser, which is sort of a thing that solves this, and it would find, given the text string, then it would output the tree and it was all beautiful. But then instead of making up a toy dictionary, I plugged it into a real dictionary like Webster's and it said, well, ‘saw,’ yeah, it can be a verb, but it can also be a noun. And in fact, every word in that sentence could be a noun because ‘I’ and ‘a’ are letters of the alphabet. And so the whole thing could be a noun phrase and I was saying like, wow, that's weird. All right, let me try an easier one. Let me replace ‘saw’ with, ‘see’: I see a bird. How could this be hard? Well, it turns out see could be the Holy See. So it is also could be a noun. I didn't know what the Holy See was, but you know, right now I do.

KEN: 12:14 Okay. All right. So I said, gee, this is crazy. It's not that these things - I mean the Holy See is possible. All right, a and I, those are possible, but they're not very likely. So then I started measuring the statistics. So we had this thing called the Brown Corpus, which was done by some friends of my parents, but is very famous now. And it was one of the first collections of a balanced corpus. They got a bunch of, I don't know if you know the local paper in Providence. And then I think they had some of the New York Times and then they had, oh, religious stuff. And they had, JFK was president and they had stuff like that. And they had all, they had all kinds of stuff and it was a million words, which was a lot in those days. And you could go count and they actually labeled every word as to what part of speech it was.

KEN: 13:03 So you can label how often is 'a' a noun, how often is, you know, 'I' a noun, how often is ‘see’ like the Holy See, okay. And in a million words, I think it didn't come up. And then you know, ‘saw,’ sure there's some cases where it's a noun but not very many or whatever. And if you replaced the sort of possibilities in Webster's dictionary with the probabilities coming out of the Brown Corpus, then all this stuff became a lot easier. And so I was giving this talk about here's a really simple solution to a problem that actually the stuff we do doesn't work that well. And I kind of expected this to be heresy because you know, when I was at MIT using statistics like this was frowned upon, you know, extremely, because - we could say why, but I want to believe that it was because Chomsky was rebelling against his professor, Zellig Harris. It's just personal, right?

KEN: 14:01 Anyway, I kind of was afraid I’d get booed off the stage, but in fact it started a movement, right? Then 10 years later it was the only thing we were doing. And then I wrote this paper about the pendulum swung too far because I worry that we're no longer teaching the stuff that I used to learn then. So, oh, I have this story, but some of the hype about deep nets is out of control. I mean, there's legitimate reasons to be excited by what's going on. Okay. And the stuff that's happening is amazing and I can go through that. But we usually, after I give a bunch of these, you know, really exciting demos, then I say, you know, but we need to be careful about what we can and can't do. And what we used to talk a lot about and what the Chomsky hierarchy is a lot about is what we can and can't do.

KEN: 14:49 And so he would say, n-grams can't do this. And Minsky would say, and nets can't do that. All right. And this is all seen as negativity. And so we're all dismissing this and we're all using the methods that they were arguing against, right. And it's working for us and people saying why if it's working, why are you worried about it? But I think it's still worthwhile to think about what we can and can't do. Maybe I'm just showing my age. Anyway. We have this thing, it's very powerful. It can do all kinds of stuff, but it can't do everything. And it would be useful to understand what it can and can't do. And so I've been sort of on a mission to try to do that. Anyway. I probably ...

CRAIG: 15:25 No, no, that's fascinating.

MUSIC:

CRAIG So from the statistical model that you introduced and that became popular ...

KEN: 15:40 There were a bunch of us.

CRAIG: 15:42 Okay. But with the validation of neural nets and deep learning, they apply a statistical model?

KEN: 15:51 They sort of start with that and then they go way beyond where we were. A lot of what we were dealing with were, you know, simple sort of linear models and these things are not linear. They're much more powerful. So one of the ideas is that if you insert lots of hidden layers, then you can do more than you could do without those hidden layers. So you hear about deep and the deep, is the hidden layers, right? They can do more. We know that, but we don't know what they can't do. Right? So the computers are better than we are now at chess, okay, no question about it. Now they're pretty good at Go. They're pretty good at a lot of these games. But for a lot of human tasks, they're not as good as people now.

KEN: 16:32 But the measurements are suggesting they are. And there are a lot of tasks where the difference between what the computer's doing, the difference between what a person's doing requires better measurements in order to see what's going on. Okay. In chess, you don't need better measurements. Okay. The machines are better, I'll concede that. Right? But there's a lot of these other tasks like, say, speech recognition where the machines are getting better, but they're - everyone knows who's used any of these things that they're not as good as a person. Okay. Right. At least a person with normal hearing and who knows the language, right? But the tests don't always show that and then people can claim that the test is right even though you know the answer. Okay. Now what's going on here? Let me go back to the part of speech tagging. So part of speech tagging got stuck soon after that paper that I wrote, there were a lot of papers on it, but where they ran into trouble was the measurements they were using were unable to show improvements.

KEN: 17:34 So the standard method is that you go and you label a bunch of text with some judges and you get the right answer. There's this thing called the Penn Treebank, which is sort of the gold standard everyone uses. Then you build a program and you try to get your program to take the same text and you're going to measure it as to whether the parts of speech agree with what's in the standard. And everybody thinks that part of speech is such an easy problem. There should be no room for debate. But it turns out there's a lot of room for debate, right? The difference of opinion, the inter-annotator agreement - disagreement rate or agreement rate, there's about 3% room for disagreement in part of speech tagging. And the machines are getting error rates that are about also 3%. There's only a difference when two people disagree.

KEN: 18:23 If I look at it, it's a difference of opinion. When a machine is different from the judge, the machine is wrong. All right? And the methods we use to measure this don't distinguish between a difference of opinion and an error, right? And, and that I think is what's going on with a lot of this. So there's a lot of claims out there that say that we're doing as well as people when in fact all that's going on is the measurements fail to see the difference, right? And a failure to find fallacy is - something anybody would know in statistics - you fail to see a difference. So you can reject the null hypothesis. I can say that this is better than that. But to say this is as good as that, that's much harder. Now let's get into measurement. Measurements now get into things like generalizations.

KEN: 19:06 So where you're going, is generalization. So one of the things that we used to do, there's a lot of cheats you can do and one is to use the same data set for both training and testing. So training is where you fit the parameters to your model. And testing is where you try to observe how well the model predicts new stuff. Generalization would be when you keep these things separate and you want to test on something different. And what's really kind of important is that, what a lot of us do is we keep separate training and test sets, but they aren't really separate. So if I'm going to train on the New York Times and test on the New York Times, that's a lot easier than if I train on the New York Times and I test on the Guardian. Yeah. Okay. 'Cause the Guardian and the New York Times, you know we do - the English is different on different sides of the pond.

KEN: 19:55 You know, you have your house style, they have their house style. But even more difficult is imagine I train on the New York Times and I'm going to test on PubMed abstracts. You know that well. You don't write about the same things that they write about in a ...

CRAIG: 20:09 Yeah. And so that generalization does not really exist.

KEN: 20:14 Well, well of course there's limits to it always. Okay. Now normally when they talk about it in machine learning, they really talk about where testing and training, where the two datasets are drawn from the same population. So I'll take some New York Times from here and take some New York Times from there, separate them and test and train. But even then, it would be different if I took the stories you wrote when Obama was president and I test on the stories when Trump's president. I mean there's words that appear on the front page of the New York Times now that you would never have used before.

KEN: 20:47 So in fact, even training and testing the news changes over time, right? And generalizing over time is troubling. And generalizing over, you know, the authors or the house style is hard, and then generalizing over subject matter is even harder. So a lot of times when people talk about generalization in machine learning, they're really just talking about how the difference between training and test. But they try to make it so the training and tests are sort of comparable. Whereas in fact, if we take a real product and we put it out there in the field, we don't control what the customers are going to do. So what you want, would be something representative of what they're gonna do. But you don't know what that is.

CRAIG: 21:35 Do you see a day when that generalization - I mean is that a solvable problem?

KEN: 21:40 I think that we can do better and better and we're going to do amazingly well. If the bar is right now, let's say, that you want to do as well as people or there some other cases where I want to do better. Well, let me talk about a case where I want to do better. So let's talk about machine translation and my lexicography friends like to distinguish what they call general vocabulary from technical terminology. General vocabulary would be the words that any speaker, the language would be expected to know and that's the stuff that goes into a dictionary. And technical terminology would be the stuff that you would see discussed in PubMed abstracts, stuff you'd see discussed in a technical conference. But this is stuff you generally can't find in the dictionary and generally isn't really general vocabulary. It's not stuff that everyone's expected to know. It's only experts in some area, but everybody's an expert in something. So the languages consist of a combination of these two things.

KEN: 22:41 Now, professional translators are really very good at what I'll call the easier vocabulary and the easy grammar, the stuff that everybody knows. And what they live in fear of is technical terminology. Because you see what happens if you get the technical terminology wrong, the experts, both the readers and the writers of these documents, know the technical terminology. And if you get the terminology wrong, it sounds like you don't know what you're talking about because the translators don't. All right? right, right. And so, Where I see the machines actually having an unfair advantage is on the technical terminology, because they could actually read all the technical material in all these fields and it's imaginable the machines could be way better than people, even the best, even the pros at that stuff. All right. And where the machines struggle is on the easy vocabulary and easy grammar, the stuff that everybody knows.

KEN: 23:37 And so what I'd like to see is a better together story, where right now I think on spelling, we're prepared to believe that machines are better at spelling than any of us. All right? But I don't think we would say the machine is better at writing the New York Times. Okay. And so there are things where I like to see more synergy between man machine interface or say, so that we would do what we're good at and then give the rest to the machine. So I do a lot of work with people in China and their English is great, but when we go socializing in a restaurant, that's when we use the technology. So they were trying to tell me about some vegetables and they don't know the English word for okra. Okay. And a lot of my colleagues who are native speakers of Chinese, but they got their PhDs here, they don't know the technical terminology.

KEN: 24:30 They couldn't give their job talk in their first language because they don't know the terminology in their native language. So that's where I think the machines could be better than people at that stuff. They can be better at spelling, they can be better at translating the hard words. Right? Yeah. But when it comes the, you know, words that you learned when they were growing up, they struggle. That's hard for a machine. A lot of times I think we're afraid of how they're going to displace us, when in fact, the right thing to think about is how they empower us. When you're talking to somebody on the phone, do you think about the phone or do you think about the somebody at the other end. Okay. When the phone's working, you're thinking about the somebody at the other end. All right. All right. But when it's not working and then you think about the phone. So I don't want to say it should be like that. It should be so good at what it does that it's hardly noticed and it would facilitate communication, not displace it.

CRAIG That’s it for this week’s podcast. I want to thank Ken for his time. For those of you who want to go into greater depth about the things we talked about today, you can find a transcript of this show in the program notes along with a link to our Eye on AI newsletters. Let us know whether you find the podcast interesting or useful and whether you have any suggestions about how we can improve.

The singularity may not be near, but AI is about to change your world. Pay attention.