AI LEADERS PODCAST – NATURAL LANGUAGE PROCESSING

VIDEO TRANSCRIPT

RYAN MCDONALD: AI can really shorten the amount of time it takes for a customer to have their issue dealt with but effectively dealt with.

FERNANDO LUCINI: So hello and welcome to this AI Podcast. I'm Fernando Lucini. I'm the Chief Data Scientist of Accenture, and I'm joined today by Ryan McDonald. And, Ryan, great time to do an intro and tell us a bit more about yourself?

RYAN MCDONALD: Oh, great. Well, thanks for having me. So, yeah, I am Ryan. I'm the Chief Scientist at ASAPP. I have about 20 years of experience in Natural Language Processing and machine learning, something that's been a passion of mine, my career for many, many years, kind of runs the gamut, worked on consumer products, enterprise products, Semantics, Syntax, machine translation and have you? Yes, so I'm very excited to be here now.

FERNANDO LUCINI: Awesome. It's always super dangerous when you put two NLP geeks in a podcast. It's one of my favorite moments because and you and I had some chance to before, but I've spent, gosh, from since 1999 working on the problem of doing something useful with what we say and what we write. So it's going to be awesome to given your length of tenure on this, to attack this problem. And what I think will probably be a great start, we're going to talk a lot about one of the specific areas where we're understanding what people say, what they mean is critical, which is the call center and such settings. And I think would be super useful for the listeners to hear a little bit more about how do you chunk this into problems within your world? What is the anatomy of the call center? So we can then start overlaying on that how some of these clever AI techniques actually try to solve it and how far they go, right? Does that make sense, Ryan?

RYAN MCDONALD: Yeah, absolutely. Sounds good. So, of course, at a call center, there's many reasons why someone might call in a call center, sales, customer care, what have you? Let me focus a bit on the customer care situation because I think that sort of reflects a lot of the different technologies that are needed to push things through. So this is the situation where you have a product or some service and you have a problem with it and you want to call up and help solve that problem.

So the first thing is, as we all know, we're usually met with some form of automation. So this can be an IVR, an interactive voice response or a chat bot, depending on whether you're calling on your phone or are doing some kind of chat going digital. And the purpose of these systems primarily is to just understand why you're calling in, right, to understand what is your problem, root you correctly to the right agents, the right set of agents that can solve that problem, gather as much information as possible, so that when you know your call gets to that agent, it can be handled effectively. But of course, in the past 5 to 10 years, really more like the past 5 years, in particular on digital, we've seen chat bots become more sophisticated, and the job is not necessarily to just route your call, but also to try to solve your

problem as well.

And this can be giving you some information, some self-service information like go here, go to this website or read this document to work through your issue or execute something on your behalf, so they might have integrations with the systems that the agents also have access to, and they might be able to execute it automatically.

But, of course, most calls still end up with agents. So it's a human agent that's going to be handling your issue. And in particular, this is true for, I would say, non-standard or non-trivial issues where a human has to look at it and actually make some decisions. And even there, AI is going to play a huge role. And this can be things like suggesting to the agent what flow to follow, what knowledge base articles to look at, even what to say next and what to do next, what tools to open and all these things because the machine learning models, they're trained on a huge amount of data, on calls that have come through the systems in the past. And they have sort of a good view of given any state of a call what happens next.

And so, we kind of make this distinction between what I'd call independent automation. So that's something like a bot. The bot is just executing by itself versus directed automation or augmentation where a human is directing still what is happening or the system is giving suggestions to the human on how to efficiently handle those issues.

And then even when your call is done, so when you hang up or the chat ends, AI is still playing a big role. This can be anything from disposition notes, so when a call is done, agents have to sort of fill out a lot of information. You know, why did that person call in? How is the problem resolved? What actions were taken on behalf of the customer? And that's still very manual. But now we're seeing a lot of systems start to do that automatically or at least assist the agent in that task.

And then, of course, there's supervisors and there's call center leaders and they are tasked with things like agent training, also analytics, trying to understand what's working at the call center and what's not working at the call center. And so, AI and machine learning play a big role in sort of looking at the calls that have come in and trying to sort of pull out the right patterns and make the right suggestions to supervisors and call center leaders, so that they can optimize their call center accordingly.

And I'm sure I'm missing things, of course, but from the moment your call comes in, to many days after your call, AI is sort of in the process all the way.

FERNANDO LUCINI: And I think that's a fascinating point of this because if you think about it from a consumer perspective, you go in there, your experience is sometimes we just the machine or the chat bot or the agent, depending on where you go. And we forget about in the anatomy of how what happens in the background, all the things that have been affected by ML or AI, beyond that what we most famously see which is the chat bot that might be helping us on having a conversation.

And you and I talked a little bit before about things like some of these magnificent things that are happening in NLP land with these very large data models, very large statistical models that LGBT 3s and all these wonderful things that give us a sense of how to deal with language and with context a lot better because they understand the general specs of language a bit better.

But in the call center, of course, this is a piece of the puzzle and it's not done all for optimization, which I think is a good segue into automation. And I think automation is being fairly - I don't know if you agree, but it's been somewhat misused in the last few years. We had the whole RPA side, the robotic automation where which is incredibly valuable and super clever, but its job is to basically, follow a macro and follow steps and just get things done that us, as human

beings, maybe don't. We shouldn't be doing because, frankly, it's just repetitive and boring, but the machine can do very well versus the automation as it relates to AI and actual AI.

So the kind of things that artificial intelligence should do very well because it actually is designed to do so and it gets an outcome better than a human being might do or the reality of an AI doing some of the automation, so a human being can do their job, which they do a bit better. I know it sounds convoluted, but you want to decrypt a little the automation because I think we have to touch on and certainly in the context of the call center. Do you want to just touch on that a little bit?

RYAN MCDONALD: Yeah, absolutely. I mean certainly, the one thing that AI is really good at is optimizing multiple variables for complex objectives, things that humans certainly have a hard time doing at any moment in time. But as you mentioned, something like RPA, they're just trying to automate like very simple things that are repetitive. Whereas, I think the real win for artificial intelligence is going to be the fact that there's all these key performance indicators that companies care about, like this might be handle time, throughput, customer satisfaction scores, things like that. And optimizing for these is challenging. How do I know when I tell an agent to say this next or take this action next that that's going to lead or maybe in the short term, that's kind of the optimal thing to do. But in the holistic view of the conversation in solving the customer's need, maybe there's a better path to take and really whether that's in the bot land or in the augmentation land, really trying to think about this holistically and optimizing for these key performance indicators. And I really think that's - I mean that's the power of AI.

And as you mentioned, we're in this great time in Natural Language Processing, models like GPT3 make what seemed impossible a few years ago, tangible at this point in time. But I think it's exactly what you said. It's a piece of the puzzle. It's a really important piece of technology that's going to help us get there and, in particular, that it really tells us how to model context and what fluent language looks like and things like that.

But we also have to take into account that the agents themselves, they're not just in the conversation. They're doing a bunch of other tasks as they're having that conversation with you. And how do we bring all that into GPT3 world or whatever your favorite massive language model is? But also, how do you optimize those models for the outcomes that you care about, like so that customers are happy, agents are satisfied and that's usually correlated. You know, when customers are happy, agents are happy, but also the businesses, the call centers are optimized for the performance indicators they care about.

FERNANDO LUCINI: Yeah, because I think one of the things that's worth decrypting is you and I are experts in the area. So I know we talk about AI as if it's a single thing, which obviously you and I know it's not. But one of the things we should decrypt a little bit for the use cases that we're talking about here is exactly that is, hey, we're talking about AI. But it's these are software programs that have to attack certain problems in this, in the flow that we talked at the beginning. So there's a piece of technology or a bit of programing that needs to deal with the chat bot area or the having a conversation area or the connecting that to content area. So how does that work? How does it work for you guys and how do you see it? So it ends up being true in the way I see it, it's almost it's a profession, right? You have a set of professions within any software environment, which is to create software programs we call AI, with data, that solves specific problems and you stack them and stack them and stack them in they solve more and more problems. How do you guys handle it and how do you see the move towards what I call the ecosystem of the AI? So where we end up having enough of these programs that are working together in, let's say, in your case, a contact center setting that they are all connected to each other in some rigorous way? And how do we make sure that balances is - you see what I mean? It ends up being like a very

complex recipe where if you get the salt wrong, the customer is not going to like the food because it is just salty, right, but it's 30 steps in between the salt and the mouth, right?

RYAN MCDONALD: So I totally know your point here and actually it ASAPP. We use this notion of AI native, which is meant to sort of encompass the fact that there's all these different AI services that are trying to do different things at different points in the call, both for the agent and the customer. And if you don't think about holistically, how to put them together in the right user experience, you're not going to get the right value from all these things, right? Like it's going to be too salty at the end of the day, as you say, right?

And so, that's as a philosophy, I think that's really important. And we certainly have a lot of studies that show no matter how good your AI is, if they're not sort of - if those features aren't incorporated together in the right way with the right UI and the right latency and really making the user experience, that the augmentation rates or the amount of times that an agent will actually take a suggestion you're offering them, it can drop off a cliff, right? All of these things are super important.

That's not to say that everything is designed holistically, like you have to break things down into smaller problems. So that can be something like auto compose, what should in a digital platform, what should the agent say next to the customer knowledge base, what is the article, the piece of knowledge, the frequently asked questions that is going to help that agent answer a particular question for the customer.

It’s, you know, developing these features, it makes sense to sort of separate them apart, have like metrics that you can iterate on. That's super important. It's really hard to iterate on 20 metrics that are highly dependent on each other that are running in a large production environment. So, for development, we really need to break these down. But ultimately, I think what we do is we have great customers. By customers, I mean our clients. So the people who run the call centers and we just collect data.

So we sort of break down all those AI services, optimize offline metrics or whatever metric we think is appropriate for optimizing that single thing. But ultimately, in order to get the whole pie correct, we need to see the thing in action and our customers, you know, we do AB tests, obviously. We try different things and make sure that we have the right parameters set the right ways for the whole ecosystem. But our customers are great and they let us experiment and try different things, change the UIs, really try to sort of optimize around the boundaries. And those AB tests, obviously, they're going to start looking at more holistic metrics, like how much throughput are we going to get through our call center, if we change something? How does handle time change? How does customer satisfaction change? These kinds of things.

So that's kind of the view we take. And, ultimately, we're moving to a space whereas we measure that and we get more and more data, we can then start optimizing for that ecosystem and really start to get into the space where the machine learning models, even though they might be individual components themselves, can be optimized in this, as you say, ecosystem or end-to-end fashion.

FERNANDO LUCINI: Yeah. Well, it's important because at least one of my pet hates is that humanizing of this thing as opposed to it being as you describe, at the end of the day, it's a design challenge like any other, right? We have an objective in this case. And I do think in your use case of the call centers being one of the more human applications of AI because you're talking to other humans and you're trying to help in that process and there's a set of AIs or a set of machine learning programs that are trying to help in that process. So it's in the cold face of dealing with a human. And as you look at how you look at this problem, it is and I love the way you put it, is a design challenge as well. It's here is my objective, which is dealing and the contact between a human and a human and the service

and the service and how do you optimize that and make it as best as possible for the whatever the KPIs, which begs the question.

RYAN MCDONALD: I was going to say, just as to point out, like we've had cases where on our offline metrics for our machine learning model that we're using, they go off the charts and we get super excited like, oh, this is going to be a big win for all of our clients. And then we'll do an AB test and we don't nearly see - we don't realize the gain that we thought we were going to realize. And it's exactly the point you're making is that if the AI gets this much better, but the UX stays the same or it doesn't change in order to accommodate that improvement, you're just not going to realize that gain.

And so, it's really important to just see these things as part and parcel, especially when humans are in the loop. Humans are like actively every step consuming this technology.

FERNANDO LUCINI: Yeah, and we should talk about that because I think I know it's important for you guys in how you see this. And it's importantly certainly for Accenture in how we see this, which is the role of the human in all of this, right? And so, I'd love to hear your view, and I know you've got many different ways of coming at it and you can pick whichever ones, but you've got the perspective of the agent, you've got the perspective of the customer and you've got the perspective of the supervisor, you've got the perspective of the person running the call center, which is a business at the end of the day, and it has its own metrics. And then there's AI is from what you're saying, there's AIs or machine learning programs anywhere inside that and you're trying to figure out what works and doesn't work. So how do you see the – I mean we talk about human in the loop. How does it work for you guys? How do you try to make all our balance of AI's plus humans work for you?

RYAN MCDONALD: Yeah, I mean, it's tricky. On the one hand, I think if you looked at it from afar, you would say, oh, there's three players in this, as you said, the customer, the agent and the call center. And in some sense, they have potentially conflicting goals in this situation. But I think when you really look at it closer, you realize that those goals may not be so conflicting after all, in particular, between the agent and the customer. We have a study called, CX, The Human Factor and a lot of agent happiness is derived from the tools they have at their disposal, the training they have at their disposal, so that they can, effectively handle the customer's problem. They want that customer to be happy just as the customer wants their problem solved effectively. So if AI can get to that, then we're at least going to make those two players happy.

You can sort of see the call center is saying like, ok, well, obviously they want their customers to be happy, but they also have a lot of customers on the phone that they need to move through the system as fast as possible. So, there's a bit of conflict there that putting too many calls in the bot, which is often not a satisfactory experience or maybe having agents just push customers through too quickly might not be satisfactory. And again, I think this is where AI is really going to bridge that gap because AI can really shorten the amount of time it takes for a customer to have their issue dealt with but effectively dealt with, not just superficially dealt with. And that's, of course, by just getting the agent on the right track, giving the agent the right tools at the right time in order to do that.

So I think like one of the promises of AI in this space is really that, to take what are potentially three conflicting players in a game, if you will, and really sort of make them realize that they have the same shared objective? And I think we're already seeing that like certainly, what we've noticed based on some of our experiments with customers is that by reducing handle time, we are seeing an increase in customer satisfaction. So the AI is both getting customers through faster, but also giving the agents the right tools to make sure that they get through faster, but with good outcomes for the customer when dealing with their problems.

FERNANDO LUCINI: Yeah, you don't want to have them repeat scenarios where you have that transaction. Sure. But it was unsuccessful and you're going to do it five times and you're super, super, super upset and you do have, as you will say, different angles because you want a customer, the ones, an answer to their exam question. And they want it as a reasonably fast, but also wants it to be pleasant and frictionless. And you know, we're all humans. We want that touch. You got the agent that's probably managing a lot of context, a lot of different context and having to do a lot of work, which I think is the consumer in the C, right? A lot of work in the background after that call between that call to make sure those connections are made depending on what are they doing, upsell or customer care or they're doing all of these that we do.

And one question I had for you, which I'm really interested. I know if you guys have done some studies on this is also that at least what I see as a trend between people wanting to pick up the phone, make a call, get something sorted out versus going to the vendor and having a quick chat on the keyboard. It's almost that move between people that feel very comfortable making the call, and they think that that's the most fastest way to get there versus people that want to text it and get it done that way. Do you see any? I say that because then you go into that idea of comprehension and which one gives you a higher comprehension or not? Where do you guys stand on that? What do you see?

RYAN MCDONALD: So maybe, maybe clarify the question a little bit?

FERNANDO LUCINI: So yeah, and apologies. It's Spanish, and I speak very quickly as well, I that can't help anybody. So a lot of the work I see with customers and all the things they're telling, certainly me and I don't know other colleagues in Accenture is that whilst we still want to prioritize the verbal communication through a call center as a way to interact with the customer, there's also a clear, clear trend of people who are really happy to text it, to use the text messaging as a message to do the same thing, but they can express themselves in a more crisp way. They don't have to have the human interaction and the way I personally see it is that more of us are always busy. And the human interaction is great, but I could do without it. I don't mind if it's human interaction on a text message. Do you guys see a lot of that happening and does it affect the way that the call center is managed?

RYAN MCDONALD: Yeah, I mean, absolutely. So I would say the majority of our clients still take voice calls. There's certainly when you look at the plots, the move to digital, to chat, is happening, but certainly in the amount of money that's being spent and also the volume of issues that are being sent in are going to voice calls. But certainly, that trend is changing. For some of our clients, we have objectives to make the digital platform experience exciting enough for the customer to accelerate that trend. And, of course, as that trend happens, like you says, this is going to change how AI works and the sorts of problems we have to face. I mean one thing, in particular, for digital is agents often deal with more than one issue at a time. On voice, they don't. They're talking to one person at one point, whereas on digital, they might be dealing with two and in some cases, three issues at a time. And this just increases the amount of contact switching and cognitive load that the agent is going to have at that time.

And so, AI can play an important role in trying to reduce that. But also interesting in this space is it really opens up possibilities for asynchronous communications. So I have a problem, I chat in, my meeting is starting, I'm just going to ignore my chat for a bit and after an hour, I'm going to go back on and respond to what the agent asked me. And it really behooves call centers to handle that kind of asynchronous communication. But that also means that the agent I was talking to initially that it might be a different agent every time I write a message, I chat in and talk, and there it becomes important for continuity that the new agent that picks it up, they quickly come up to speed on what the problem was and what's

been done already feamlessly as possible. And this is again like a great place for AI to come in, like to summarize, to highlight like what's happened? What was this person's issue? What has previous agents done for this person, so that when it gets to you and it's been two hours since the last communication, you can kind of answer as though you were waiting for them the whole time.

And so, I think that sort that shift has sort of opened up all sorts of like interesting challenges.

FERNANDO LUCINI: Yeah. So one of the of them and we should talk a bit about the difficulties as well. But one of the challenges I seen there, when we talk about GPT3, when you've got all the other wonderful big data transformers for this, there's things like summarization, for example, there's things like that that you would have thought that you sit there and there's an agent that sits there and talks to a new customer. And there's this wonderful narrative that comes out two paragraphs that they can read quickly that says this customer came and they took this and they did that and they were super happy. But then we failed them in that or whatever. And the truth is that while that sounds fabulous, I'm not quite sure where they're in terms of doing those kind of things. So to be good to talk a little bit about the limitations where AI can go within your context and where we'd like it to go and where it's not a classic one, I guess we can start with this is simply the understanding of voice. I mean it's moved incredibly from large dictionaries to neural nets to, I think, now everybody's is now moving into transformers even for this, which is the latest and greatest fad.

I shouldn't call it a fad, but fantastic and wonderful and magnificent, magnificent evolution. Now, from your experience giving that you're on the sharp end of this with accents, 16 kilohertz versus eight kilohertz telephony and all this stuff that makes it difficult to function. Where do you think you are in terms of some of these technologies getting you where you needed to be and what are the difficulties you think?

RYAN MCDONALD: Yeah, so certainly, I think you've hit the nail on the head in terms of ASR quality. I mean it's fantastic. And I think fantastic compared to where it was a few years ago, and if you get native speakers speaking clearly with no background noise, not interrupting each other, everything like that, then the transcripts are quite good. And we actually have an amazing world class speech team. And I'm shocked sometimes at the quality of the transcripts that are coming out. But obviously that is an ideal world and people have accents. People aren't native speakers. People are walking on the streets on their cell phones and busses and ambulances are coming by and the transcripts can be quite poor in many cases when this happens, and because I would say all the AI that follows, whether that's trying to understand sentiment, trying to make predictions to the agent about what to do next, trying to write summaries about what happened during the call. These things all rely on relatively clean transcripts and they're going to suffer. You're just going to get this compounding effect.

And I mean this has been true of all NLP technologies since time. The amount you sort of stitch together, the more compound areas you're going to have. And so really, whether that means like end-to-end models to reduce that or just more advances in ASR, I think that's going to be important.

The other key on ASR is also latency. The models that we see now that are amazing, sometimes we need to run them on very expensive hardware. And even then, they may not have the latency requirements that we need. Things have to really operate in milliseconds in order for any kind of directed automation to be successful. Like if I suggest something for the agent to do, even if it's three seconds too late, that agent's already moved on. They've kind of made a choice and moved on. And so, those transcripts and our ability to do it real time in order to really help the agent at that moment, that's also a problem because then the less resources you have, invariably the lower the quality your transcripts are going to be.

FERNANDO LUCINI: And you guys also deal with small amount of context as well and time. And it's not like you're dealing with somebody reading a whole tale or the Bible, right? You have somebody giving you a phrase. And with that phrase, you have to fulfill, understand, transcribe. And I'll tell you what one question that I for you, which I think is always interesting for all of our customers and all the AI work, which is we've moved from creating these wonderful things that are predicated on black and white, right? And somebody can use this system because it's simple. You put everything in a database, you can see what's put in there. It's very discreet. And suddenly we move to this work of shades of gray. Was this transcript very good, not very good. The business has to live with the fact that you and I are creating AI systems where there is inherent variability on this and summarization is a great one where some of these amazing technologies can create you a summary that you think, oh my lord, this is like a human would have done this exactly as well, this is great versus the same machine. Because the data was not as good giving you a summary of you think, okay, I can't use this.

So how do you guys deal with that? And how do you manage the getting the machine learning or the AI to a useful production state without basically not doing anything because not everything is perfect? How do you tend to manage that?

RYAN MCDONALD: Yeah, absolutely. And I think this goes back to really one of the first things we talked about in the conversation is getting the user experience correct. We need to put safety nets around these things, so that when they do fail, when they do go off the rails, I mean summarization is a great example, like text generation there's been a slew of studies in the past couple of years on the fact that they hallucinate, that they can make up facts, especially when the source material isn't reliable. How do you deal with that? And, of course, one way to do it is say like, well, I'm going to have some confidence threshold that the system has and I'm going to tune that. And I'm only going to show people the output of the system when the confidence passes that threshold. And that's certainly, I think, a way to do it. But I think the best way to do it is just make that user experience as good as possible. And what I mean by that is give the agents the tools so that they can quickly assess is that prediction something I can use or not. Should I say this? Should I not say it? Here's a summary at the end of the call, I can try to fully automate that, so that the agent's not needed, but another way to do that is let the agent quickly look at my notes and give them the appropriate UI so that they can select them, edit them, remove them quickly. So we're still saving that agent a lot of time, but we're giving them sort of the tools that they need to make sure that only high quality information is passing through.

And that's why I think getting back to this like AI native concept, making sure that these advanced machine learning technologies are coupled with the right sort of safety nets through user experience to make sure only the good stuff is getting through.

FERNANDO LUCINI: Yeah, and I've never heard of it referred to as hallucination, but I will be stealing that. That is a wonderful thing, an AI that's hallucinating. That it's the most popular thing because they do and is the best-known geeky way of putting it.

We've spent a lot - this has been an amazing session. We've gone through a lot of depth here. It's like if I were to summarize, I think what I've heard today in closing, I think it's we're hearing that we need to design AIs, like everything else, we have to design them into a journey that has all the right flavors for the consumer, for the agent and for whoever else is involved in that. It is all about that design, making sure that it can actually do its job in the right way. And I think it's also clear that it's all of the service of these nice humans that are trying to get either a service or give a service. And so, the service of these folks and how they interact with that is super important that more and more, I think, we're going to start seeing all of these interconnected

systems or software or AIs, which whatever way you want to think about it and actually thinking of them together as a little net of, as you say, native AI, is a sensible way to start thinking, given that it's going to become much more part of your life and you're going to deal with it. And that, hey, AIs hallucinate. Not often I hope, but sometimes they hallucinate and we just need to put the right sensible governance around them, so they can be doing their best. And when it's not their best, it's okay because it's built into the journey.

So that was an amazing conversation, Ryan. I'd love to thank you very much for taking the time to have this wonderful conversation. And obviously, thank you to the listeners for geeking out with us and, hopefully, they're not hallucinating too much. I'm going to be hallucinating a sentence. So with that, I'd like to ask everybody that's listening, if you like what you're hearing, you enjoy it, please, please subscribe and share with your friends and your colleagues.

And once again, Ryan, thank you so much for taking the time to chat with us. And no doubt you and I will geek out again in the few days and weeks coming. Thank you very much.

RYAN MCDONALD: Well, great. Thanks for having me. I really enjoyed it.