EPISODE 61

[INTRODUCTION]

[0:00:10.6] SC: Hello and welcome to another episode of TWiML Talk, the podcast where I interview interesting people doing interesting things in machine learning and artificial intelligence. I'm your host, Sam Charrington.

The show you're about to hear is part of a series recorded at the Georgian Partners Portfolio Conference last week in Toronto. My guest for this interview is Kenneth Conroy, VP of Data Science at Vancouver-based Finn.ai; a company building a chatbot system for banks.

Kenneth and I spoke about how Finn.ai built its core conversational platform. We spoke in-depth about the requirements and challenges of conversational applications and how and why they transitioned off of a commercial chatbot platform, in their case API.ai, and built their own custom platform based on deep learning, word2vec and other natural language understanding technologies.

Georgian Partners is a venture capital firm whose investment thesis is that certain tech trends change every aspect of a software business over time, including business goals, product plans, people and skills, technology platforms, pricing and packaging.

Georgian invest in those company's best position to take advantage of these trends, and then works closely with those companies to develop and execute the strategies necessary to make it happen. Applied AI is one of the trends they're investing in as our conversational business and security first. Georgian sponsored this series and we thank them for their support.

To learn more about Georgian, visit twimlai.com/Georgian, where you'll also be able to download white papers on their principles of applied AI in conversational business.

Before we jump in, if you're in New York City on October 30th and 31st, we hope you'll join us at the NYU Future Labs AI Summit and happy hour. As you may remember, we attended the inaugural summit back in April. The fall event features more great speakers including Corinna Cortes, Head of Research at Google New York; Davide Venturelli, Science Operations Manager at NASA Ames Quantum AI Lab; and Dennis Mortensen, CEO and Founder of startup X.ai. For

the event homepage, visit aisummit2017.futurelabs.nyc. For 25% off tickets, use code TWiML25. For details on the happy hour, visit our events page at twimlail.com/events.

Now, on to the show.

[INTERVIEW]

[0:02:46.4] SC: All right, everyone. I am here at the Georgian Partners Portfolio Conference in Toronto, Canada where I'll be doing a few interviews. I am excited to be here with Kenneth Conroy. Kenneth is VP of Data Science at Finn.ai. Kenneth, welcome to This Week in Machine Learning and AI.

[0:03:05.7] KC: Thanks for having me.

[0:03:07.1] SC: Awesome. You're in from Vancouver.

[0:03:11.1] KC: Yeah. Just in from Vancouver from last night, so feeling at a bit, but happy to be here.

[0:03:17.8] SC: Nice. Why don't we get started by having you tell us a little bit about your background and how you got involved in Data Science and machine learning?

[0:03:23.4] KC: Sure. Well, I'm from Ireland and I did my PhD in Dublin City University in the area of heterogeneous sensor data. I was in the research lab called Clarity Center for Sensor Web Technologies. I was involved with coming up with a way of aggregating sensor content from multiple different devices that were not in any way connected.

In doing that, I ended up coming up with machine learning strategies for clustering. I have one for supervised learning for finding out things that researchers wanted to detect. A research is in this case would be domain experts in sports science.

From there, when I finished my PhD I continued working in the same university but for The Insight Centre for Data Analytics. That's where I moved way into the NLP machine learning sentiment analysis projects. It was a commercialization wing of the university essentially.

I ended up moving to Vancouver maybe three years ago. Now I'm at Finn.ai and leading the data science team as we create conversational assistants for banks.

[0:04:29.2] SC: Nice, nice. Is there a lot of machine learning and AI activity in Ireland?

[0:04:32.5] KC: Yes. There's a very big club actually in Dublin. I'm half hoping that we set up a Dublin branch as for our [inaudible 0:04:38.1] headquarters in the future. I know a lot of people in the field and the universities, very qualified people there. Yeah, it's a bit of a hub right now for machine learning.

[0:04:48.4] SC: Nice. So tell us a little bit more about what Finn is up to.

[0:04:51.7] KC: We're creating conversational assistants for banks. Think of a Siri for your bank. We write things like day-to-day banking, your basic needs like checking your balance online, or paying bills, or transferring money from friends to friends, as well as more kind of day-to-day Q&A types of you would ask a bank; what your fees are for your credit cards, product information, stuff like that.

We integrate with the bank's APIs for security and logging in, if we have any interactions with your bank accounts. Essentially yeah, we're just expanding our product base beyond the initial NLU side of things for detecting intents, to having recommendation engines in the future, as well as credit score coach, moving more into the voice platforms.

Right now, we're NLP-based, text interface mostly. We have experimented with some voice interfaces, but our primary platform is Facebook, for which we launched a production bot for ATB financial as of yesterday.

[0:05:49.6] SC: Okay. Congrats.

[0:05:50.3] KC: We're in production as of yesterday.

[0:05:52.7] SC: Nice. Nice. You're here at the conference as a portfolio company of Georgian Partners, which you're also speaking at the conference later on today. What's the topic of your talk?

[0:06:05.6] KC: The talk topic title went through some iterations, so I think how we built a virtual banking assistant, or virtual financial assistant for banks.

[0:06:15.5] SC: Okay. It sounds like that's the Finn story and –

[0:06:17.8] KC: Essentially yeah. One of our co-founders Natalie Cartwright, she'll be introducing the talk and giving a lot of the business insight into why we went down the path we went down and then I'll talk more to the data science side of things and how the technical challenges arose along the way and how we make the decisions we made and how we're continually improving on that process.

[0:06:39.4] SC: Okay. Well let's dig into that. What were some of the major goals and challenges on the data science side?

[0:06:46.7] KC: I think we reduced the problem to its core issue, which was how do you design a taxonomy of intents for which you want to get answers? It's a bit of a complex task, because you've got to – for us, we were a closed domain. We wanted to stay a closed domain and stay in retrieval-based systems.

As banking only and their responses are pre-written, no generative responses at all. When we have that as our canvas and we want to create a taxonomy to drill down into individual branches of what conversational train might be, there were a lot of challenges involved in that where you draw the boundaries between things.

At what stage or to what extent do you want to make it conversational with one-to-one responses? The bulk of our functionality in terms of intents would be one level. We would say a question and you get an answer. But some of them will have flows, so if you want to make a bill payment you can say something like, "I want to pay my BC hydro bill $50." It will pre-fill those things using engineering recognition and ask you to confirm. Or if you don't give enough details we'll ask you for more input. That to and fro conversational side of things comes into it.

[0:07:57.9] SC: In that case, it's almost like walking through a wizard in a traditional UI where you've got a set of questions and represent information that's needed to fulfill the transaction and you just go on to the next, to the next.

[0:08:09.3] KC: Exactly. Based on conversations we have with customer service agents, so a lot of our functionality is to help the CSAs answer questions. In doing that, we got a list of questions that they get on daily basis and we could wait them based on how frequent they came in.

From that, it was like the seed of how we build this taxonomy, that one-to-one communication with banks. As we speak to more and more people and if different banks were seeing the commonality between the banks is very strong. There is maybe 85%, 90% commonality in what they want to achieve.

That's the good side of it, having this taxonomy in place you can – we'd strap a bot fairly quickly and get it up to a reasonable level of performance very quickly. It's then when you build out those more of fully featured functions and those conversational side of things. That's when complexities come into and where the future is for the business.

[0:09:10.4] SC: A bunch of questions. I guess one question is not necessarily related to the data science side of things, but are you integrating one-to-one with the banks, or are you using some kind of intermediary like Yodlee or something like that?

[0:09:24.2] KC: I think, so I don't know too much about the – what goes on behind the scenes of the engineering side of things. The engineering team will do a lot of the integrations with the banks themselves.

A lot of our bots are completely done by us, hosted by us and accessible by anybody, because they're not personally – no personally identifiable information in there. Well for banks that require API integrations with the banks themselves, we provided an interface for the banks to give the security credential requirements and login functionality. A lot of stuff is done by them. We do not take the responsibility of authenticating the users. That's up to the banks to do if they want to have integration for their systems.

[0:10:05.2] SC: What does it mean to have a banking bot that doesn't deal with any personally identifiable information? Are you saying that you're just passing requests to the bank and they're answering them? Or are you talking about bots that just don't deal with account information that are telling you where to find branches and ATMs and that kind of thing?

[0:10:24.8] KC: Yeah. Similar functionality would be finding ATMs, or just basic questions about the products and services they offer. That stuff is not confidential in any way. It's open information that may exist on an FAQ and a different format on their side.

It's still a conversational way of talking to your bank or e-bank. It doesn't necessarily have your bank if you're not authenticated with it. But yeah, that's what I mean.

[0:10:49.5] SC: Okay. What percentage of the banks you're working with are taking this approach where you know it's just that surface level interaction versus account level detail?

[0:10:59.7] KC: Typically when we do PLCs, so four banks. The first version will not have banking integrations. Banks move a lot slower than startups for instance. If they want to integrate their systems with us, we have had a lot of trust between us. So we've had a relationship with ATB Financial for over a year now. Based on that level of trust and that level of working collaboratively together, we've been able to create a production bot that uses all of that API integration.

Yeah, building up that trust and building up system for them and allowing them to test it and kind of iterate on it before it goes to production is key.

[0:11:39.0] SC: Okay. We started talking about the data science side of things in terms of identifying the intents. In what way can you characterize the breadth of the intents or the interactions that you have for a typical customer? How many of these intents are there in a realm like banking?

[0:12:06.9] KC: This question come up a lot. What is an intent? We use a lot of entry recognition to reduce the amount of intents. We do not want a lot of intents in our system. We want to break down the problem and taking a smaller subsets, more consumable subsets.

The amount of intents is actually quite low compared to what you would think based on the responses we can give. We give thousands of responses, but there might only be a few hundred intents. It depends on the agent, the functionality that we have it depends on how many FAQ-type questions they want in there and what subset of all functionality that we can offer they want to make use of.

Yeah, the amount of intents is a wishy-washy number. The amount of responses is a solid number that we can give on a per agent basis. But the amount of potential intents is set at a certain value for our core basic banking needs. So the taxonomy tree as an empty list of responses. Anywhere you could put a response in any of those leads and that's another response.

That's up for the banks to decide what they want to support. The banks come to us with their call logs information analytics on what the subject matter is for their child logs for instance. Then they can see a subset of the taxonomy that that would best meet their needs.

[0:13:27.4] SC: What level of abstraction is an intent? Is it something like I want to ask a question about an FAQ, or is it like individual questions? Does it map more closely to individual FAQ questions or responses?

[0:13:42.1] KC: It depends. Things like if you want to find an ATM, would have like a find ATM intent. It's fairly well defined. If you want to find out your fees for your Mastercard for instance, the intent might be fees. The entry recognition would find what the Mastercard is or means and provide a different response based on Mastercad rather than Visa. That's the way we break it down into smaller bite-sized chunks.

[0:14:05.3] SC: Interesting. You've got this tree structure, you've got these entities. I guess kind of basic question, are you using any of the commercial bot platforms to do any of this stuff, like the API.ai **[0:14:21.8]** and others?

[0:14:24.2] KC: Not anymore. We did.

[0:14:26.5] SC: Interesting.

[0:14:27.0] KC: Initially we used API.ai way back when. They're great for what they're built to do. They allow you to create bots fairly quickly. Not much data involved. Accuracy isn't too bad, but fairly customizable. But they weren't the way that we could continue. They were too limited in terms of what we wanted to achieve. We had no idea what was in their model for instance. We had no secondary match for what their primary intent match would be.

[0:14:53.3] SC: You said no idea what was in their model, meaning they identified intents using NLP whatever, but you know is all a black box to you. They just gave you an API and told you, "Hey, bot user initiated this intent and we found these entities and figured out, but you couldn't tweak or tune that at all."

[0:15:15.9] KC: Exactly yeah. There was that limitations on where we could go with it if we want to have multi-layer models for instance or focus on – or wait specific types of intents. The functionality just wasn't there to be in a scalable enterprise product within our domain. Because we were in our domain, we wanted to have a system where we could reuse as much of what we were providing for each bank across banks.

API.ai is quite complex when it comes to sharing data from one agent to another and you need individual agents for that. There's also a limitation on the amount of data you can upload to it and the rate limitation on how often you can request model matches and so on. The other reason why we moved away and that was the move or approach in system was a 100% it's our own now.

[0:16:05.4] SC: You mentioned in there a multi-layer models. What would that entail, or what does that represent?

[0:16:12.5] KC: In place of our entity recognition, we are building up data sets to find how the user goes down a certain tree. Right now we're using anti-recognition to do that, but we could use a model to do that, like a subset of all available intents, but narrowed down to those individual entities. Once the data set has been more richer and a bit more well-defined and developed than we have more user data, we can use that as an additional layer in the model.

The existing model, which is the top layer intent match and then within each conversational flow there are sub models that have all available actions there, as well as a few escape actions. So rather than having to cancel out of that flow to query something else, you get a focused way of finding what the user's next input is.

[0:17:01.0] SC: I'm still trying to wrap my head around what the – so the idea of multi-layer, meaning you've got your entities that you're extracting out traditionally, but then you're applying

other probabilistic models and try to identify where the entities are, is that the idea behind multi-layer?

[0:17:18.1] KC: The multi-layer model isn't in production, but this is a replacement for part of that and to your recognition modules. Given enough data and enough information on what the synonyms and aliases are for different entities in our system, we can further define and automate that process rather than using the existing one, because you got to update things like sentinels over time, and you might miss certain things.

You might not always find Mastercard if it's misspelled for instance. But that said, both that approach and the existing approach if you misspell Mastercard, because we're using the intent level at a higher level, we're still finding it what the intent of the user is, which is say fees, and we could just guide the user via UX to this specific response they're actually looking for.

[0:18:07.0] SC: Okay. Is part of that shift to the multi-layer, is there an underlying shift and approach from something analogous to NLTKM Python to something that's more like a homegrown deep learning entity recognizer?

[0:18:24.2] KC: Yeah. So we have two tracks right now. We have our production track, which is it's a spec Python pipeline, using word embeddings and using word2vec or other algorithms like that. Then we have our research R&D track. In that track, we're working on deep learning stuff. We're working on MXNet and tensorflow in parallel. We're experimenting where we're going to get the best accuracy over time and where this can fit in to our multi-layer approach.

[0:18:57.4] KC: Tensorflow is very popular and MXNet is very good.

[0:19:05.6] SC: I have heard MXNet is very fast. I don't know that I've heard very good necessarily. Not that it's not, but –

[0:19:13.1] KC: I can't speak with too much confidence on the subject.

[0:19:15.1] SC: Or maybe define good.

[0:19:17.1] KC: I can't.

[0:19:19.1] SC: Okay. Got it. Interesting. I mean, there are tons of – lots of framework choices out there.

[0:19:24.5] KC: There are a lot there. Yeah, it's early days for our deep learning path, but the goal was to get this out into production and a usable state and have our stakeholders happy with the performance of it, which they are, which is great. But this is the exciting future side. It's getting everyone seem very excited working on the deep learning side of things.

[0:19:45.7] SC: Nice. Nice. Are you able to characterize in some way the benefit that you – you know besides from the flexibility and visibility, agility, all that ility kind of things, like in going from API.ai where you didn't control the model to that first pipeline that you just mentioned where you're using word2vec and stuff like that? Was there a specific quantitative advantages that you saw?

[0:20:15.5] KC: Yeah. When we're evaluating, we couldn't evaluate API.ai in the same way. That we didn't the ability to see, split the training test data on their side and find out what the cross foundation scope would be for instance. We can do that and objectively evaluate our own models against their own models, which is great.

But we do analyze logs and we take a snapshot of everything that's gone through our system as it exists in API.ai and then put it into our new model as well. Then compare the results of that in real terms. We take a snapshot of weeks' worth of data that goes into an agent and see how accurate it is based on that.

That's what the banks are interested to see too. They're interested to see the real-world accuracy of something. We don't adjust thing for duplication in the data or however cross-hold violation. It's hard for them to visualize how that's working. So real-world, they need to see what those values are. In doing that, we can see that our models are outperforming what was existing for API.ai.

[0:21:15.7] SC: Okay. We started down the path of challenges. Are there other things that come to mind in terms of challenges that you've come across?

[0:21:25.1] KC: I guess. I mean, we haven't seen it too much, but it can happen and we're quite aware and we're kind of getting ahead of the curve like cautioning against it, is intent drift. When

the banks themselves have the ability to map utterances to intents, we are at risk of having our taxonomy start to diverge based on how they write the response and how they think that should be the answer for that specific intent.

If you got a bunch of data in there that's a label of an intent and all those pieces are answered by that response, then part of that response gets taken away. That doesn't make sense for those things that were in there initially to still map to the same place. It look like it's doing the right thing, but the responses would be given to the user.

So to counteract that, we are building tools for the banks to be able to do this themselves and to CSAs. With our human in the loop infrastructure, we'll be able to correct errors made and map things to the correct intent and then feed that back into our system.

We also have methods of coming up with approximations of what those things are, kind of prompting the user to make from some subset of those things. But we want to have a two or three level deep verification on anything they map to make sure that it is appropriate response to what the user's utterance was.

[0:22:48.7] SC: Can you give me an example of where they might change something that disrupts that relationship?

[0:22:55.4] KC: Sure. We have things like tell me about you and who are you. Sometimes the response to that give us the same thing, but they're separate intents.

[0:23:05.2] SC: Tell me about you and who are you.

[0:23:08.7] KC: Tell me about you would be like, what can you do? Whereas, who are you is just an introduction to the bot itself. But sometimes who are you means who are you with the bank? So what train that goes in for that is important depending on the context, what the response is.

[0:23:28.7] SC: You mentioned that the CSAs, it sounds like to some extent the customer service teams at the bank have some role in defining out the hierarchy and the responses. Is that correct? Or are you doing this all as a service for them?

[0:23:53.4] KC: Earlier collaborations, very heavily involved with meeting with the bank's CSAs, talking through their issues, how their data forms their opinions on what should be in there or whatnot. But we do have a base taxonomy based on those consultations with multiple banks at this point? We can see about 85%, 90% of commonality across all the banks.

That's our core taxonomy. If banks want to extend beyond that, they just know that they'll have – we'll have fewer data entries for those things, few utterances to train. It would be a little less accurate at first. But if it makes sense, we can build that in as part of our core offering. In other cases, it might be locale-based or might just be for that one bank. In those cases, they're just one offs kind of customizations.

[0:24:40.0] SC: Okay. So this case where you get drift in the intent. Is that because of changes that the customer service people are able to make in the underlying system, or is it the training data that they feed to you, that you are putting through your pipeline?

[0:25:00.7] KC: The customer, which is the bank has the ability to change the responses if they want to change the responses. We kind of monitor that to make sure it doesn't go too far away from the meaning of what that intent is.

But they're also able to use our systems to manually answer questions. If a query comes in and we do not have the high enough confidence to give the answer, it gets handed off to our human in the loop system. Then at that point, they can answer the question and then recommend an intent or utterance to map to that.

Over time as well, they can also monitor their own logs and if they have people who are qualified to do so and can educate them to do so. They can mark those logs for us, for themselves. Then that could be fed back into their model.

This happens a lot in pre-production when we have user accept in testing with our banks. They pre-train the bot by interacting with it time and time again. Typically we annotate that, but we want to hand over some of that to the banks too to give them an insight and to the type of things that are being asked how we best answer those questions.

[0:26:03.5] SC: Okay. It sounds like the intent drift challenges almost the skill/cultural/focus kind of thing. How much of that is addressed by educating the customer service folks versus technology solution?

[0:26:25.8] KC: It's a bit of both. It's not just the customer service folks too, it's anybody in the organization. They might say they want a thousand questions answered, but that is 1,000 distinct intents in any way, or even sometimes their exact same intent, because that same subset of intent, so I think part of it is to frame the problem differently.

People still tend to think of things like FAQs and a chatbox if it's an FAQ on a webpage. But those have the questions don't get asked in the same way as it would be listed on a webpage. You got to structure things differently. It's not conversational, it's not intuitive, and the user doesn't just do it.

Part of it is managing the expectations from the start and training the users, customers on how we best organize this information on why we do that. As well as, because we've got the taxonomy that's kind of shared across banks we can reuse the data across banks. That's mutually beneficial for everyone involved, because it means the accuracy of all of those intents improve as data from each of those banks come in.

[0:27:32.8] SC: Interesting. In your presentation, will you be outlining – do you have any prescriptive – if you're looking at doing chatbox or maybe even ML or AI generally, like do things one through end?

[0:27:48.9] KC: Kind of. Yeah. It's obviously heavily weighted in the way we actually did it and how we tried it and how it didn't work and then how we did it and it seems to be working pretty well. So yeah, I'll talk about exactly how we did it, as well as how we're best seeing rolling this out works to customers. So seeing this in the real world, seeing how customers interact with it and help you get to where you need to go and where themselves need to go, it's really cool to see.

[0:28:17.3] SC: What are your top three make sure you do this, or don't make this mistake kind of advice to startups that are trying to get systems like this up and running?

[0:28:27.9] KC: Don't rush into it. Make sure the taxonomy is well defined. You will spend a lot of time researching this at the start. Two, is you got to get some good data in there. We looked out lots of sources of information, including previous chat logs or transcripts from calls. They're not great. Even if I see that they're not very good and they're often unstructured and difficult even, annotate in any way. We found for the cold start problem crowd sourcing is key. Getting a list of different ways of saying different things can get you very far very quickly.

[0:29:05.0] SC: Did you crowd source among the customer service agents at Customers or like Amazon Turk or something like that?

[0:29:11.6] KC: Yeah, both.

[0:29:13.1] SC: Okay. So interesting.

[0:29:14.0] KC: Mechanical Turk for things that customers weren't helping build. Maybe their features for our core platform that have not been released yet. Then often, customers would have their own ideas and they would submit those ideas. We would build on those by submitting Mechanical Turk jobs to get more data.

I guess then, the third thing would be to consider where you draw the line between intents and entities, how you reduce the problem space and increase the accuracy. By customizing to a certain extent to the domain you're working in. So get that taxonomy up and running, get that dictionary of terms for banking up and running if you're in banking. Obviously, I don't like to do a banking book.

[0:30:01.0] SC: But you said a bunch of stuff in there, so the first thing was the line between the entities and the intents, and we haven't talked about that yet. Is there like a tradeoff dial in there somewhere that you have to find the right place?

[0:30:16.8] KC: It's not a straightforward task to find out where to do that, or what the top intention be. Like do you focus on a product and then have a bunch of sub entities that have – from features of that product, are you talking about the features of products that correspond to that feature, or a thing like fees, whatever.

The way you segment that stuff, you got to be able to keep track of what you've done as well if you ever want to go back again and change things. You don't want to be doing that when you've got a few million data entries in there to annotate. It's going to be an impossible task.

That's an issue with a lot of the first time bot builder. You just build a bot and it works and you're like, "Okay, cool. It works." Then you want to add another 10 or 20 questions, then you got a lot of conflicts because there's lots of similarity between those questions and existing questions. That's the downside to doing that approach, and the upside to having a vertical and a well-defined taxonomy to take into account everything that could be asked in that vertical.

[0:31:17.3] SC: Is there a right or wrong way to approach building out that taxonomy in terms of this line between the intents in the entities, or is it more you just have to do something and stick to it and be like absolutely consistent as you expand out your vocabulary?

[0:31:38.7] KC: Yeah. I mean, there's no right answer to this. It's a very difficult problem to solve. It's not that you can't change it afterwards, but you want to change it a little bit. You want to tweak it, you might split some intents or combine other intents. But you don't want to completely rebuild the system every few weeks or months. So yeah, it's a bit of both.

[0:31:57.4] SC: But there was another question that I wanted to ask. It's maybe going a little bit back to – you scrapped API.ai and built your own system for doing this, and you said it was Spark and Python. Did you also invest in building almost like a user interface or a higher-level platform tools for both of your internal people and the banks? Or how sophisticated does that layer need to be in order to have a usable system?

[0:32:35.8] KC: Yeah, it's a very complex system and it's I guess 70% built right now. Most of the internal functionality, we don't have the external bank facing side of things that we're going to get. But it's essentially our interface for our database. It's where define what intents belongs to each customer, what the taxonomy is for that for instance, what entities are in there, where they apply, what the response content is and the ability to map from utterance to those intents, and support the user in making those decisions too.

So to make sure that you got the context as to what that intent means. If there are a few hundred intents, it might be difficult to know exactly what an utterance should go into if you're

not fluent in the system. We have some supporting information, meta data associated with that or examples that already exist to help you make that decision. Yeah, that's called the Atlas Project and that's our – I guess, the heart behind everything.

[0:33:32.2] SC: Interesting. Interesting. Then in terms of we were speaking upstairs earlier and you mentioned that you as part of getting to this initial production customer, you kind of re-platformed everything. How did you have things deployed before and how are they deployed now?

[0:33:51.3] KC: Before, we had API.ai for doing a lot of our functionality for enter the recognition for instance, as well as the model itself. But now, we have that entirely ours. It's a rebuilt from –

[0:34:04.4] SC: Okay. I guess I was wondering in terms of – I guess, now that you've built your own platform to replace API.ai, what processes and automation have you built up around deploying models out into production?

[0:34:20.8] KC: Yeah. The offline stuff, so like the training of the model, we don't have well-defined release processes for that side of things. But any model we build and we release, we have a strict release process that's done with engineering and dev ops. We ensure that we have a record of what model is deployed in which environment, what it was trained on what the results of the evaluation were over time.

It's controlled. We don't just release models whenever we feel like it. We also have test waits that we must run through before we ever release anything to production, to make sure we've not broken anything. If any intent has reduced in accuracy and that affects something else down the line in terms of maybe a flow or a conversational flow or some keyword isn't working in some point in the flow, we got to make sure that's caught and not deployed.

I think we're maturing as a company and that we're a lot more strict on our deployment environments and how we're actually doing things. But yeah, everything is all setup on the cloud and got a serialized model to produce the intent recognition side of things. Yeah, it's working pretty well.

[0:35:32.6] SC: You mentioned that you're also able to catch when intents change meaning and things like that in testing. But are you also able to – do you have machinery in place that is able

to identify – I guess, I'm thinking of it like a long tail of intents or entities that aren't being caught by the system. How automated does that need to be? Is that something where you run a weekly or monthly report and just look at what isn't being caught correctly, or is there an automated process in place that is always up to date and you're always trying to catch up to kind of build out that long tail?

[0:36:15.7] KC: Yeah. I think to some extent, some of the long tail were never going to go down too.

[0:36:20.4] SC: It's a long tail.

[0:36:21.3] KC: It's a long tail, sure. We do have the human in loop functionality. It works really well for our used case. When it comes to finding the things that we probably should support, oftentimes we will have the ability to run it through our actual model, our model trained on everything that we have, rather than the model of things that we support for that agent.

Then we can find out if there are things that were accurately catching, but they just don't have a response for. So it looks like we just don't support it. That catches some of those things. We can also use and supervise clustering to find groups of things or questions that converge around a new topic, or a new question type that we may want to suggest to the customer to include, or for us to include in our taxonomy if it is not already.

We also do manual log analysis. For the time being, we're keeping an eye on the data. A lot of the data is especially now we've got a production data coming in. It's interesting to see firsthand, help you make decisions on how we refer to the product from a customer point of view.

[0:37:27.4] SC: Interesting. Interesting. It sounds like you guys are doing some really interesting things, and in particular it's interesting to hear some of the background behind the platform shift and what's enabled you to do. So thanks so much for sharing that, and I'm looking forward to catching your talk.

[0:37:45.2] KC: Awesome. My pleasure.

[0:37:46.9] SC: Awesome.

[END OF INTERVIEW]

[0:37:52.3] SC: All right everyone, that's our show for today. Thanks so much for listening and for your continued feedback and support. For more information on Kenneth or any of the topics covered in this episode, head on over to twimlai.com/talk/61. To follow along with the Georgian Partner series, visit twimlai.com/gppc2017.

Of course, you can send along feedback or questions via Twitter @twimlai, or @samcharrington, or leave a comment on the show notes page.

Thanks once again to Georgian Partners for their sponsorship over the show. Be sure to check out their white papers, which you can find by visiting twimlai.com/Georgian.

Thanks again for listening and catch you next time.

[END]