Bruke Kifle:

This is ACM Bytecast, a podcast series from the Association for Computing Machinery, the world's largest education and scientific computing society. We talk to researchers, practitioners, and innovators who are at the intersection of computing research and practice. They share their experiences, the lessons they've learned, and their own visions for the future of computing. I'm your host, Brooke Kifle.

As technology rapidly evolves, NLP stands at the forefront from GPT to Gemini and Lama, language technologies or large language models, as they're known, are reshaping our intelligent systems at a rapid pace and transforming how we generate and interact with information. Amidst this transformation, however, it's crucial to equally prioritize the vital importance of inclusivity and language technologies, considering their substantial impact on access to information and opportunities. Our next guest, Partha Talukdar, is driving advancements to machine learning and NLP while advocating and ensuring for more inclusive and equitable language technologies.

Partha is a senior staff research scientist at Google Research India, where he leads a group focused on natural language processing. He's also an associate professor at the Indian Institute of Science Bangalore. Previously, Partha was a postdoc fellow in the machine learning department at Carnegie Mellon University. He received his PhD in computer Information Science from the University of Pennsylvania. Partha is broadly interested in natural language processing, machine learning, and in making language technologies more inclusive. Partha is a recipient of several awards, including an Outstanding paper award at ACL 2019, and the ACM India Early Career Award 2022. He's a co-author of a book on graph-based semi-supervised learning. Partha, welcome to Bytecast.

Partha Talukdar:

Hi, Bruke, great to be here, and thanks for having me.

Bruke Kifle:

You have such a remarkable and interesting career that spans both academia and industry. Having your undergraduate experience in India, coming to the US for your graduate studies and postdoc, and then now returning. I'm very interested to learn, what are some of the key points within that personal and professional career and journey that have led you into the field of computing, but also motivated you to pursue your field of study now with language technologies?

Partha Talukdar:

Sure, yeah, happy to chat about that. Yeah, so I got into computer science. I had some exposure to computer science during school days, but it was really during the undergrad where I took it up as my major and especially transitioning into NLP and AI. That really happened during an internship, that summer fellowship that I got at Indian City of Science where I actually have a faculty position now during the third year of my undergrad. So I think it was about 2002 to summer basically. I was working on networking technologies before then, but when I got that summer fellowship at IIC, which is what Institute of Science is shortened to, so I got exposure to language processing. It seemed like really interesting, and I just, one thing led to another. 20 plus years after that I'm still working in language technologies now.

I was really fascinated at that time in terms of how we can extract information from languages. Interestingly, I was still working on low-resource languages and then veered off into doing other things in NLP, and now back in Google research, I'm back into working on languages with limited resources, how we can make it more inclusive. So it has been a full circle that way, both in terms of geography and also topics within NLP that I covered in my research career so far.

Bruke Kifle:

And I think it must be interesting, clearly now if you ask anybody about ChatGPT or GenAI, I think it's the hottest topic of the year, but I'm sure 20 years ago where when you initially made your journey into this field, it was a pretty new domain and area. Now we have these LLMs that are generating a ton of buzz, but clearly over the course of a decade, two decades, there's been a lot of transformation that we've been seeing in the AI space primarily as a result of deep learning. So what are your thoughts on language as a pathway to achieving artificial general intelligence, which of course is the north star of the goal? Mainly when we think about language as a very pivotal role in human cognition and communication. How have some of the advancements that you've seen in LLP really led you to think that there might be something here?

Partha Talukdar:

Right, yeah. Language as a central component, as you rightly said, in terms of communication and cognition is super important. And recognizing this right from 1950s, there has been the work on language processing. In fact, people early on thought that machine translation will be a solved problem in say 50s and 60s, but of course it took much, much longer than that. While we have made significant progress, but still there is still even within translation, there is a lot more work to be done. So that way I think the importance of language processing and NLP has been there all along. But of course with language modeling, in fact large language modeling, making all the progress in the recent years have really brought it to Limelight. In terms of, say how I have seen area transition during my research career so far, so earlier it was for say, different tasks even within NLP, let's say if you're interested in information extraction or machine translation or say parsing. So you would have customized models for each one of those tasks separately.

And then even within NLP, it was quite challenging to not change topics because there is a lot more groundwork that you have to do in order to build the baseline systems in the particular subtopic that you're working on. So things like that has seen a sea change now with one, like say, general purpose, pre-trained language model that you could either use with some instructions or with some fine-tuning, you could make it do multiple tasks, not only within language processing, but increasingly with multimodal systems, you are able to go across modalities as well. So the same model working with say, speech, images, text. So all of those would've been very hard to predict back in the day that within this short span of time would come to this homogenization in terms of the modeling.

And I think that has happened in stages. So initially with neural networks and deep learning, that there was homogenization. First with machine learning, if you go all the way back from rule-based to machine learning, so we started working with data-driven methods. So in that case, maybe the algorithms were the same, but you were building different models and also maybe different types of algorithms like say SVMs and decision trees or CRFs and all. Now with neural networks and deep learning, there was standardization in terms of the learning algorithms. So now you use say, deep learning for all of it, but for different tasks you would still use different models. Now with language models, now there is homogenization in terms of model also. So now you have in a single model doing multiple things. So that way in stages there has been more and more standardization and homogenization across tasks and across modalities, which has enabled for researchers to move across these different problems. And then also sharing of and transfer of knowledge across these different tasks and learning.

So that way, lots of changes. And then going back to your question on the importance of language, if we look at all the advances that we have been celebrating, large language models have been at the forefront and have demonstrated the way of how these self-supervised learning done at scale could result in lots of interesting capabilities that would've been hard to anticipate a few years back.

Bruke Kifle:

You raised a very good point around this idea of generalizability, where we have single pre-trained foundational models that are basically able to serve multiple functions or multiple use cases. But I would presume a big part of that is this generalizability, the ability to learn things, not just for specific tasks, but to really capture intelligence, I guess depending on how you define intelligence. But do you believe that with these LLMs we're seeing a case of systematic learning or as some schools have thought, are these just giant databases? Are they actually exhibiting some of the learning capabilities, the reasoning capabilities that are basically essential to deliver on a lot of the tasks, whether it be summarization, sentiment analysis, Q&A, what are your thoughts?

Partha Talukdar:

So I think definitely there is an acquisition of the knowledge of the world, but in addition to that, I think there is definitely, it's not just memorization and reproducing what is there in the data, ability to understand intent of what the users are saying, and with those limited instructions, trying to fulfill a task requires lots of reasoning and generalization capabilities. So I think there is no doubt about that. Even if you say ask to write a piece of text in the form of, say, some author who may have never written about in that particular style. So trying to say write about, say LLMs in the style of the Shakespeare. So these say language models will happily comply with that type of request, but nowhere in our say, pre-training or fine-tuning data, we would have those combinations being present.

So that way it's able to learn some patterns and apply that in novel situations, which was clearly not present in the training data. So that way there is definitely generalization happening, but to what scale and how those things are happening or I guess in the topics of ongoing investigation, I guess we have very little understanding on that as a community currently.

Bruke Kifle:

Yeah, very interesting. One thing that your research area or focus or passion is on is around making language technologies more inclusive and accessible for low resource languages or underrepresented communities. And of course, when you think about how these technologies are used in the real world, in real products, there are serious implications, whether it be digital divide or the unequal access to information or opportunities, there are serious implications when we don't make these technologies inclusive. So could you describe some of the current gaps or limitations when we say NLP in the context of low resource or underrepresented languages, and then what are some of the ongoing efforts to try and address some of these gaps?

Partha Talukdar:

Sure. Yeah, no, I'm very happy to talk about that and that's in a topic I'm really passionate about. But Brooke, before going there, I just want to mention about the importance of language about your first question. So would it be okay to address one more point there?

Bruke Kifle:

Yes, Please.

Partha Talukdar:

Okay. Yeah. On that particular point, one of the things I guess which has resulted in this excitement is the ease of interface between these AI models and humans, which has happened through, again, a natural language. So now I think all of these capabilities, just by giving some instructions right now or some problems, people without having any expertise in computer science or machine learning are able to deal with these language models and have access to AI capabilities. I think that has really increased the scope and has excited even non-experts to and make those capabilities available to a broader mass. So to me, natural language as an interface I think has been, and recognition of that, I think has been a major advance through this language model. So that's one part. And then the other one is how these language models have acquired all of these say the knowledge about the world, which has resulted in this impressive capabilities.

That's also true large corpus where NASA humans over the years have documented about not the knowledge about the world. So both in terms of storage of knowledge of the world around us, and then also as an interface language, interface like to these AI models, I think again, language has been a core enabler in both of these two aspects. So now coming to your data point about inclusivity. So I think the question is that now if we know, say English and a handful of other languages, then the capabilities of these AI and language models are available to us. But there are vast majority of languages in the world where these models don't do very well. So just to give you some idea. So currently these language models may work well for say a few dozen languages, or even if you look at say language technologies more broadly, even beyond say language models, we have capabilities in maybe a few dozen languages across the world, but in the world we have more than 7,000 languages there.

And for vast majority of those languages, basically beyond these in a few dozen, we have no usable language technologies, be it in the form of speech technologies, translation technologies. And so that way it runs the risk of those who have access to these technologies, the access to information and opportunities will become more and more easier while living behind vast part of the world's population from having meaningful opportunities to leverage these technologies. So that seems not an ideal state to be in. So my research and my group's work has been focused on how we can make these capabilities available to speakers of a larger number of languages.

So now when I think of LLMs, so the notion of a yin and yang comes to my mind. So as I mentioned that say if we leave things as is, then this gap between depending on the languages that you know, the access and opportunities to information that you have depending on a known language, I think that's going to grow. But at the same time, I feel that language models are also the best tools that we have at our disposal right now to reduce this language-based barrier that's there, but that's going to require some concentrated effort on our part to make these capabilities and models more inclusive. And that's where our research has been focused on. So one, I guess thing is that when we are looking at this broader set of languages, we are talking about a diverse geographies, people with coming from different cultures.

So one thing that we need to understand is, what are their core needs, what language technologies they need that could best serve their use cases? That's, I think one part. Then say if we look at say, scaling the existing capabilities, be it say translation or speech recognition, synthesis to speakers of more languages where it makes sense. So lack of data is a challenge. So right now the recipes that we have is that if you have data and compute, we have good methods to build these models, which can have a very interesting capabilities that we all have seen, but for lots of languages around the world, they don't have as good a representation on the web or a lot of them, the representation on the web, they're not well represented. So that creates a challenge.

So how we can work with communities, with speakers around these different geographies and cultures to make technologies relevant to them while dealing with these data sparsity problems is a central issue as we look at scaling these methods. And at Google we have this thousand languages moonshot where we are looking at building language technologies for a thousand languages around the world, and we are trying to address some of these issues in a systematic manner.

Bruke Kifle:

I see. So the data issue is definitely a big problem. This idea of data scarcity, you mentioned somewhere on the order of 7,000 languages, maybe a small select few that are probably well served by these existing models. How do we address this data scarcity challenge at scale? What are sustainable, scalable solutions?

Partha Talukdar:

Right. Yes, that's a super important question. One is that the importance of representative data. When we are trying to build these models, it's important that whatever the end use cases and the communities that's going to use these types of models, it's important that all the nuances that are there, which is going to affect the user experience when they're using these models so that those parts are covered. Now, one example of that effort is an effort called Vaani that Google is supporting an Indian initiative of science is driving that program. So there the goal is to collect representative speech data or in fact, as we call it, collect the speech landscape of all of India. So we are collecting image prompted speech data from all districts of India, and it's motivated by the fact that language, when it's spoken, there is a variation across regions, even one language in a multilingual society, depending on what other languages are being spoken, there is variations in how that language gets spoken.

So in Project Vaani, that's why we are taking a region anchored approach rather than a language anchored approach where we show people locally relevant images in a particular district. A district is like say county in US, and then we ask users or contributors to describe those images in a language of their choice. And we are really amazed that when you give people this opportunity to express themselves in a language of their choice, rather than being prescriptive about it, the diversity of languages that they use to describe, say in this particular case, the images. So we have had instances in Vaani where people use an endangered tribal languages, and we've never thought about collecting data from all of those languages. So that way, capturing all of these diverse data, and then we are also making all of this data open source, and then 10% of that is being transcribed.

So building these data ecosystems where multiple organizations and even people who are passionate about building language technologies, how we can pull in all of our resources to collect representative data, which covers the underground variations, I think is super important, and Vaani is one effort of that kind. Second one is when we are, say building these models and then deploying them and making them available across cultures and geographies, there are variations in terms of local norms, the axis of harms that are there in these different geographies. For example, if I take a case of India from a responsible AI lens and then contrast that with say US. So while we have shared axis of discrimination, let's say for gender, but in India we have additional ones, let's say caste or regionality. So it's important that we test our models in these region and culture specific dimensions and take a necessary mitigation steps.

So to make sure that these models have been tested and mitigated from these fairness and bias perspectives, which we call as the recontextualized REI. So basically the responsible AI which have been recontextualized in that target geographies and culture perspective. So it's really important that we work with communities who have been historically at the receiving end of these societal biases that we work with them, understand their needs and what kind of discriminations are there. Because otherwise it might be very hard to predict what kind of issues are there if we don't have lived experiences and on the ground knowledge about these things. So two points here primarily. So one is, working with representative data and then also finding scalable ways of working with communities to make sure that the responsible AI aspects are covered. So for that, we have an effort called Bindi where we are trying to do exactly that using a complimentary approach of doing scaling, using LLMs, and then also working with communities to learn from their rich and varied experiences and then incorporate those into the models.

Bruke Kifle:

I think that's such a great point. Beyond the data representation issue, which leads to the language or usefulness of some of these models for certain communities, I think even understanding the idea of region specific biases I think is a very, very important point. And you raised a really good example because I think over the course of maybe the past five or so years, there's been a strong community around responsible AI, and I think a lot of the fairness principles, the ethics principles that we explored for classical machine learning are now very different in the age of LLMs, right? We're thinking about new harms and new responsible AI concerns. But I think the two solutions, one being rooted around representational data, but then two, close collaboration with communities to ensure that most cases we may not even be aware or understand the local or regional context. So I think that's a very, very great point that you raised.

ACM Bytecast is available on Apple Podcasts, Google Podcasts, Podbean, Spotify, Stitcher, and TuneIn. If you're enjoying this episode, please subscribe and leave us a review on your favorite platform.

One thing I want to touch on, you lead a group focused on NLP at Google Research India, and we talked about some of the work you shared, some of the work that's happening around the inclusive workstream. What are some of the key other projects or initiatives your team is currently working on that you're excited about?

Partha Talukdar:

Yeah. So a lot of our work or pretty much all of our work is centered around large language models and how we can make them inclusive and responsible. So I mentioned about the thousand languages moonshot at Google. So a lot of the work that we do is part of that initiative. So we also, since we are situated in India, we think of India as a microcosm of the global south and try to take inspiration from here and try to develop methods with the hope that if we are able to build something that works here that could be applicable more broadly in other geographies and locales with similar characteristics. And we have had some success in that direction and that we want to do more. So inclusion and responsible LLMs, linguistically inclusive LLMs and doing that in a responsible way has been our core focus. And initially we started off with text as the modality, but now we are also increasing that to include speech as an additional modality and going more towards multimodal versions of these kind of models.

Specifically in the Indian context, one of the problems we are looking at is how we can build language models for a hundred plus Indian languages. India is a very linguistically diverse country. So we have based on last census, which happened in 2011, we have total some 1300 plus languages, and we have about 120 plus languages, which are spoken by a hundred thousand plus speakers, each and 60 plus languages, which are spoken by more than a million speakers each. And then in the constitution, 22 languages are officially recognized. And language technologies right now are available maybe around for, say the top 10 of these languages in terms of say, number of speakers. So that way, as you can see, there is a big gap in terms of the large number of speakers of these languages for whom either there is no usable or no language technologies at all. So we are looking at how we can build a speech-text model for speakers of all of these languages.

Bruke Kifle:

I think that's very exciting. And you raised a really good point, which is India has a large population and not just in numbers, but also in diversity, linguistically as you've mentioned, culturally. And so being able to develop solutions that are able to cater to such a wide and diverse group can actually serve as a great model for scalable solutions that also serve the broader global population as well. So I think it's quite exciting to be able to not only innovate and work on advancements in this space, but being able to do it in a context or a region or a population where you're able to get the same level of diversity that you're able to achieve when deploying these kinds of solutions to larger groups. So that's very exciting.

One thing that I do want to call out is along with your co-authors, you did receive the outstanding paper award for your work around word sense disambiguation, which of course is topically related to some of the work that's happening in LLMs. Can you describe the problem for those who maybe are unfamiliar and maybe just share what the key insight or set of insights of your paper were?

Partha Talukdar:

Sure, yeah. So word sense disambiguation is the problem that in many languages, including in English, the same word could take different meaning depending on the context in which that particular word is being used. For example, if you take the word bank, it could mean a financial bank. So I went to the bank to deposit a cheque versus say river bank. So say I took a nice stroll along the bank. So that word sense disambiguation problem is given a particular context. How do you identify which sense a word in a particular context like now, which sense is it basically expressing? So that's the word sense disambiguation problem. So usually you have some pre-identified senses of the words and the word sense disambiguation problem is that how you can develop a algorithm or a method which given a word in a particular context can tell you out of say the end possible sense possibilities, which one is being expressed there?

So people had looked at this problem as a problem of classification with discrete labels. So given the words, say bank in a particular sentence, which one of these say four senses is it being expressing here? So the key idea in the ACL 2019 paper was to think about those senses not as like, say discreet labels, but think in terms of their embeddings. For example, if we could represent them as a vector in some vector space, then we could then expand to new and unseen senses that were not seen during training data. So why this was important was because when you are treating these senses as like said, discrete labels for classification, so the senses that you had not seen in your training data, there was no possibility of predicting those senses during test time. So basically if you have an unseen sense that shows up during test time, you will have no hope of making that prediction.

And then also for the words, some senses are more popular than the others, and then the popular senses tend to get a more biased treatment by learning algorithms. So we had some embedding based methods to overcome some of these problems, and we also showed how lexical knowledge in the form of say a word net, where you have word to word relationships in terms of whether one word is antonym or synonym, or you have say glosses, which are these short examples and definitions of the senses. So utilizing those other supplementary resources and knowledge, we basically demonstrated a way of making this word sense disambiguation problem more robust and flexible and extendable to new senses that may not have been seen during training time.

Bruke Kifle:

I see. And when I think of use cases, I think specifically machine translation, I've primarily observed many cases where I've seen this issue that you're describing where the wrong word or the wrong context is used. Is this a use case where this technique can bring some improvements in quality?

Partha Talukdar:

Yes. Yeah, definitely. Let's say a great example. And then also this work was done in a pre-LLM world, and now with another language models do a very good job in terms of learning the meaning of words in a contextual manner. But I think another possibility of utilizing these other lexical knowledges like say word net and all how we can still incorporate them in a language model is also an interesting question.

Bruke Kifle:

Very interesting. Okay. On the topic of papers or publications, I also want to quickly touch on a book that you authored a while back, which was on this idea of graph-based semi-supervised learning, which is a combination of two areas, semi-supervised learning, and then graph-based learning. At the time, what were some of the challenges that you observed in semi-supervised learning or opportunities that led you to think about graph-based approaches? And then now in this world of deep learning, LLM, generative AI, are there practical examples or use cases that you see in the context of NLP?

Partha Talukdar:

Sure. Yeah, absolutely. So I guess now all throughout my research career, so sparsity of data has been a common theme, and now we talked at length in terms of linguistic inclusion and not the data sparsity problems there. But even before that, when I was looking at say, more information extraction or how we can bring more knowledge about the world into machine learning algorithms. So again, lack of data was a recurring problem. So think of if you're interested in learning about various types of entities and relationships, be it like say people, mountains, diseases, islands across the world, what relationships are among them? So if you're thinking of doing this in a supervised learning setup where you provide training data for each and every type of knowledge, since there are so many different types of knowledge, you cannot provide lots of label examples for all of them.

So you can only provide maybe a few examples of say people or a capital of countries and so on and so forth for different types of relations. So the problem I looked at was how we can, given some small number of examples for large and diverse types of knowledge, we can build machine learning models. So this is again in a pre LLM era I'm talking about. So the observation was that doing all of these labelings through humans was a time-consuming process. So we can only get access to a small number of these learning examples, but unlabeled data is available plentiful. And what I mean by that is say corpus on the web or documents is available plentiful. So how we could utilize those unlabeled data with small amount of labeled instances to combine them to learn in a good, say learning models to say, extract, classify whatever end tasks we were interested in.

So that was the motivation for semi supervised learning. So where you combine small amounts of label data with lots of unlabeled data. Then graph came into the picture. Graph is a very useful and versatile data structure. We are all connected in one way or another. So networks and graphs provide you a very flexible way to represent knowledge about the work, be it let's say one person connected with or related with another person or an institution in a social network, or be it a biological network or transportation network or knowledge graphs. So that's knowledge about the world and relationships where the nodes are represented as entities and edges represent relationships among those entities. So that provided a flexible way of representing various domains and world knowledge. So that's the representation part. And then how we can do learning over those type of graphs with limited supervision is how the semi-supervised learning part came about.

So how we could combine these two pieces. So that's where the graph-based semi-supervised learning came into existence. And so people had looked at utilizing graph-based learning for other problems, but some of our works and other researchers around that time were one of the first ones to apply those ideas within NLP. Some of our initial applications where, say you give maybe five examples of watch manufacturers and then given that data and say access to the web, how you could significantly expand the list of those, say watch manufacturers and extract a hundred others from the web. So given those small number of five examples, so those are some examples, and then subsequently it went into how we can build these knowledge graphs, which are these entity relationship graphs that I talked about.

So one big project that I was involved with during my post-doc time at CMU led by professor Tom Mitchell is a project called NELL, which stands for Never-Ending Language Learning, where the idea was to basically build this knowledge graphs by reading web documents in a pretty much self-supervised manner, and then rereading this knowledge and using that knowledge to improve the extractor and build this in a never-ending manner.

So for that particular project's context, it ran for about 10 years in pretty much in a self-supervised manner by basically applying these semi-supervised learning ideas in a graph context. So yeah, I think is a one concrete example of basically merging graphs and semi-supervised learning.

Bruke Kifle:

Very interesting. I think it's cool to see some of the practical applications, but also benefiting from the combination of two methods of learning, both the semi-supervised and the graph learning. So that's very awesome. I think we touched on a lot of interesting things throughout your research career. One thing that's actually quite impressive is that you wear multiple hats, right? So you, in addition to your research role at Google, you are also an associate professor at the Indian Institute of Science. So I think that's quite interesting, being able to balance both have a foot in academia, but also in industry and research. How do you balance your time and your priorities between these two roles?

Partha Talukdar:

Right. Yeah. So I'm currently on leave from university, so where I'm not teaching on a regular basis, but I continued even starting the position at Google, continued advising my students, the PhD students who have graduated now. But I think it was challenging, but since I was working roughly in the same related areas, so that way it was not dragging in different directions. Also, my students were already towards the second half of their PhD journey, so they were already quite independent. So that definitely helped in terms of managing the two sites. And then now we have this collaborative projects that I mentioned in Avani before, so that's also within university, the same university. So now the engagements have morphed into different types, but there is a strong back and forth.

Bruke Kifle:

I see. And I was going to actually touch on what are some of the benefits of working in both settings, but I think you alluded to the potential for collaborations between both academia and some of the work that's happening at Google Research. Do you find that having a role in industry helps inform some of the research that happens in academia? Is it vice versa where we have some research in academia that's helping push some of the innovations in product, at least for you, which hat do you find inspiring or informing more of your work?

Partha Talukdar:

Yeah, no, that's a great question. And in fact, I also had a startup in between, and one of the primary reasons why I'm at Google is motivated by the fact that I wanted to see whatever research I'm doing if and how that's making any the impact or use in the real world, and basically go the full path and understand how it's getting used, what are the drawbacks, and then take inspiration from there to inform the next set of research questions. So I feel that that's a very productive way of making sure that you're working on important problems. And that's why having a strong connection with industry, because many times industry is at the forefront in terms of deploying products and users using those products, the problems they get exposed to, all of those. So making them available to academic researchers and influencing them to work on that is definitely, I think, helpful. And that has been one of my motivations of say why I did the startup at Google, as I mentioned.

And even before this, during my PhD time also, I spent about a year at Google in three different internships, and that had a strong influence on my research trajectory. So that way I had an exposure towards the knowledge about the benefit of an industry engagements. And I just continue to follow that even today.

Brue Kifle:

I think that's a great point. Of course, exploratory research is equally important, and I think it's essential for pushing the boundaries of science, but also thinking about grounding some of the research we do in the context of real-world problems or real-world use cases can also be quite beneficial to ensuring tangible short-term benefit as well.

Partha Talukdar:

And also Brooke, you can always take the inspiration from those real-world use cases, and then you can think in terms of say at what timescales you want to solve them, right?

Bruke Kifle:

Yes, yes.

Partha Talukdar:

And then also, this could be an individual researcher's taste, so how much they want to be influenced by that. Sometimes you want to solve that problem exactly, but in other cases you want to keep that flavor in mind, but think about what could be a more general version of that problem and then try to address that in a more systematic way. So I think having that exposure I think is very valuable. At least that's what I have found to be valuable. And depending on individual researchers' taste, they could decide how to incorporate that in the research.

Bruke Kifle:

Yes, yes. That's a great point. You introduced something which was very cool, that you had a entrepreneurial venture as well, which was quite interesting. But it seems like over the course of your career, you've spanned diverse areas, but still maintained a central focus on language technologies on NLP. And I'm sure this past year or two years, three years has been quite exciting. I feel like the large language model advancements have been quite remarkable, but every day there seems to be something new in the news. Could you highlight maybe any of the recent breakthroughs or findings? It can be in the context of LLMs multimodal. I think we're seeing some interesting work happening in that space that you find particularly exciting or impactful?

Partha Talukdar:

Yes. So I think in terms of say, broader technological arcs, so exactly the two things that you mentioned, language models, in particular multimodal models, I think have opened up lots of interesting possibilities, both in terms of use cases and then also additional research to be done. So that I think from a technical perspective, I think it's an exciting time to be in. But of course not everything is solved as we discuss at length today in terms of how we can make it more usable and helpful for a broader set of diverse users and people from different backgrounds is I think still a very, very open problem and an exciting problem at the same time. And what really excites me is that there are enough foundational research work to be done, and if you're able to make progress on that, the possibilities of societal impact are massive. So that way as a researcher, it really excites me.

Beyond multimodal and language models more broadly, more recent advances around long context models where we are able to specify a lot more contextual as part of say a Gemini 1.5 models, also I think have opened up interesting avenues for exploration, and I'm excited to explore those.

Bruke Kifle:

Very exciting. And then when you think about, of course, your primary passion or area of interest, which is ensuring that language models or language technologies are inclusive and accessible, are there things that you see as the next key achievement for the space to ensure that these technologies are more inclusive? Is it focusing on data? Are there innovations on the modeling aspect? Are there other things that you perceive as being important regarding language representation, but also some of the fairness issues that we described in the context of LLMs?

Partha Talukdar:

Yes. So I think there are, maybe we are thinking through not three dimensions. So one is representative data that we talked about, and Avani is an example of that, the Project Bindi, which is looking at the fairness and responsible AI and the importance of working with communities while leveraging the scale that LLMs give. The third one that we haven't talked about so much is in terms of how we can do these models in a more scalable and modular manner. So right now, the recipe for building these models are, you have say one monolithic model and then you try to add more data and then try to extend its capabilities. But I'm not sure whether that's a highly scalable approach. So in order to overcome that, we've been working on a method called ACOM, which is looking at how we can maybe develop models with different expertise independently and in a post-hoc manner, still see how we can compose these models to enable new capabilities.

So with that, maybe I have a core model which is very good at doing say, reasoning tasks or math or numeric reasoning problems, but it works primarily well for English and a few other languages. But then if I want to make those capabilities available for say, Santali or Hausa, but if I have a separate model with expertise in those languages, so how we could compose both of them together to enable the reasoning capabilities in all of these additional languages. So we have had some initial promise and success in that direction, and we are excited to follow up more so even in terms of modeling how we can do this in a more scalable and modular way.

Bruke Kifle:

I see. So focusing on representational data, focusing on improvements in modeling, and then of course community-based development. So working with communities on these solutions.

Partha Talukdar:

And then looking at more broadly the recontextualized responsible AI so that we are serving the end users in a locally sensitive manner that brings meaningful change to their lives.

Bruke Kifle:

Yes, yes. I think this was a very interesting discussion. I want to wrap up with one question. As somebody who's had a very diverse career, both as a researcher, as a professor, as an entrepreneur, I'm sure you've had the chance to teach and engage with and mentor many students, many junior colleagues who have gone on to become successful, whether it be as researchers, as leaders in their own fields. What are some of the skills and qualities that you look for and try to cultivate in your students or mentees? And then what advice would you give to young aspiring engineers and scientists who want to make an impact in this world?

Partha Talukdar:

Right. Yeah. So one is making sure the focus on quality is always there, not compromising on quality and a high bar for some short-term gains, even if it requires you to stay the course for a longer period. So I think maintaining quality is, I think, one important thing. Making sure that you are passionate about another problem that you're working on, and then you actually care about the outcomes, because I think that's going to help you navigate through downturns that are bound to happen when you're working on not challenging problems. So identifying things that you really care about and making sure you are focusing on them, I think is important. Curiosity, drive, I think have been important ingredients, I think in order to identify good problems and also do good work, eventually. Importance of question, identifying the right question to address is also extremely important.

So I tend to believe that even a suboptimal answer to the right question is more valuable than an optimal answer to a suboptimal question. So spending enough time making sure that you are working on problems that you care about and are impactful is I think important. And if it makes sense, then seeing how maybe this is going to be grounded in the real world and how it may help end users, if that's a thing that you care about, is something to maybe think about early on and how it's going to fit in into the bigger picture and not just looking at what's the next incremental improvement that could be done.

Bruke Kifle:

Wow. I think those are all great pieces of advice. Identify a question you're passionate about or interested in solving. Be curious. Never compromise on quality. And where relevant, think about how your work ties into society and the larger context. So I think those are all amazing pieces of advice for the next generation of makers and creators. So with that, thank you so much Partha. I think this was a wonderful discussion, and we look forward to the many impactful work that you will continue to contribute in this growing evolving technology landscape.

Partha Talukdar:

Yep. Thanks Brooke. Great talking with you. And thanks for giving me this opportunity.

Bruke Kifle:

ACM Bytecast is a production of the association for Computing Machinery's Practitioner Board. To learn more about ACM and its activities, visit ACM.org. For more information about this and other episodes, please visit our website at learning.acm.org/B-Y-T-E-C-A-S-T. That's learning.acm.org/Bytecast.