00:03

I'm Kalika Bali, I'm a linguist by training and a technologist by profession, I have worked in academia, in startups, in small companies and multinationals for over two decades, doing research in and building language technology systems. My dream is to see technology work across the language barrier. As a researcher at Microsoft Research Labs India I work in the field of language technology and speech technology. And I worry about how can we make technology accessible to people across the board, you know, irrespective of the language that they speak.

00:43

So natural language processing, artificial intelligence, speech technology, these are very big words, they are buzzwords right now. Everybody is talking about what exactly is NLP or natural language processing. So in very simple terms, this is the part of computer science engineering that makes machines process, understand and generate natural language, which is the language that humans speak. When you are interacting with a bot trying to book your train tickets or flight tickets, when you are speaking to a voice-based digital assistant in your phone, it's natural language processing that underpins the entire technology that makes that work.

01:26

But how does this work? How does NLP work? In a very, very basic way, it's about data. So a huge amount of data of how actually humans use language is then processed by certain algorithms and techniques that make the machines learn the patterns of natural language of humans, right?

01:52

These days, another buzzword that you hear a lot about is deep neural networks. And these are the advanced techniques that underpin a lot of the NLP stuff that happens right now. And I will not go into the details of how that works, but the thing that you really have to understand and keep in mind is that all of this requires a humungous amount of data, natural language data.

02:18

If you want a speech system to converse with you in Gujarati, the first thing you require is a lot of data of Gujarati people speaking to each other in their own language.

02:33

So 2017, Microsoft came up with a speech recognition system which was able to transcribe speech into text better than a human did. And this system was trained on 200 million transcribed words. In 2018, an English-Chinese machine translation system was able to translate from English to Chinese as well as any human bilingual could. And this was trained on 18 million bilingual sentence pairs. This is a very, very exciting time in natural language processing and in technology as such. You know, we are seeing science fiction, which we had read about and watched, kind of come true in front of our own eyes. We are making giant leaps in technical advancement. But these giant leaps are limited to very few languages.

03:29

So Monojit Choudhury, who's like a very good friend of mine and a colleague, he has studied this in some detail and he has looked at resource distribution across languages in the world. And he says that these follow what is called a power-law distribution, which essentially means that there are four languages, Arabic, Chinese, English and Spanish, which have the maximum amount of resources available. There are another handful of languages which can also benefit from, you know, the resources and the technology that's available right now. But there are 90 percent of the world's languages which have no resources or very little resources available. This revolution that we are talking about has essentially bypassed 5,000 languages of the world.

04:19

Now, what this means is that resource-rich languages have technologies built for them, so researchers and technologists get attracted towards them. They build more technologies for them. They create more resources. So it's like a rich getting richer kind of a cycle. And the resource-poor languages stay poor, there's no technology for them, nobody works for them. And this divide, digital divide between languages is ever-expanding and by implication also the divide between the communities that speak these languages is expanding.

04:52

So in Microsoft, in Project Ellora, we aim to bridge this gap. We are trying to see how can we create more data by innovative methods, have more techniques to build technology without having a lot of resources, and what are the applications that can truly benefit these communities. So at the moment, this might seem very theoretical, like what is he talking about, data and techniques and technology. So let me give you a very concrete example here.

05:24

I'm a linguist at heart, I love languages, and that's what I love talking about. So let me tell you about a language that many of you might not know about. Gondi. Gondi is a South-Central Dravidian language. It is spoken by three million people in five states of India. And to put this in some kind of perspective, Norwegian is spoken by five million people and Welsh by a little under a million. So Gondi is actually a pretty robust and pretty large community of the Gond tribals in India. But by UNESCO's Atlas of Languages in Danger, Gondi is designated vulnerable status. CGNet Swara is an NGO that provides a citizen journalism portal for the Gond community by making local stories accessible through mobile phones. There's absolutely no tech support for Gondi. There is no data available for Gondi, no resources available for Gondi. So all content that is created, moderated and edited is done manually.

06:34

Now, under Project Ellora, what we did was that we brought together all the stakeholders, an NGOs like CGNet Swara, and academic institutions, like IIIT Naya Raipur, a not-for-profit children's book publisher, like Pratham Books, and most importantly, the speakers of the community. The Gond tribals themselves participated in this activity and for the first time edited and translated children’s books in Gondi. We were able to put out 200 books for the very first time in Gondi, so that the children had access to stories and books in their own language.

07:11

Another extension of this was Adivasi Radio, which was like an app that we built and developed in Microsoft Research, and then put out there, along with our stakeholders, which takes a Hindi text-to-speech system and allows it to read out news and articles provided by CGNet Swara in Gondi language. Users can now use this app to read, watch news and access any information through text and voice in their own language.

07:44

A very interesting thing is that this app is now being used to translate -- by the community to translate text from Hindi to Gondi. Now, what that will result in is a lot of parallel data, that we call parallel data, that will allow us to build machine translation systems for Gondi, which will truly open up a window for the Gond community to the world.

08:06

And what is even more important is now we know how to do this. We have the entire pipeline and we can replicate this for any language and any language community which is in a similar situation as the Gond tribals.

08:21

Also education -- yes, you know, information access -- yes, but what about earning a living? Right? What about -- how can we make these people earn a living through the digital tools that all of us just take for granted these days? Vivek Seshadri, who's another researcher at MSR, and his collaborator, Manu Chopra, they've designed a platform called Karya for providing digital microtasks to the underserved communities. His aim was basically to find a way to provide a means of dignified labor to the populations, the rural populations and the urban poor populations of this country. They don't have access to all the knowledge to use the digital platforms that all of us use every day without even thinking, right? But ... Here is a large literate population that wants to work, right, and how can we make this possible for them? So Karya is one such way through which this population can get on to the digital world and, you know, through that find work and do tasks that can then earn them money.

09:35

So we saw this and we thought, oh, this is wonderful. We could probably use this for data collection as well. So we went to Amale, which is a small village of 200 people in the Wada district of Maharashtra and decided to use Karya to collect Marathi data.

09:50

Now, I know what you are thinking -- I'm sure a lot of Marathi speakers also in the audience -- that Marathi is not a low-resource language. Marathi is definitely a mainstream language of the country. But as far as language technology is concerned, Marathi is a low-resource language.

10:06

So we went to this village and we had a very successful data-collection trip. And, you know, this village is very remote. They have no TV, they have no electricity, they have no mobile signal. You have to climb a hill and wave your phone around if you want to, you know, use your mobile to call anyone. So they gave us all this data. But more than that, they gave us very valuable lessons in life.

10:34

One is this pride in one's own language. The people of Amale were thrilled to be doing this because they were advancing their own language by doing this. The second was the value of community. Very quickly, this became a village community effort. People would gather together in tasks and do this together as a group. And the third is the importance of storytelling. People of Amale were so starved of content that in the morning, during the daytime, they would do recordings of stories in Karya and then in the evening they would gather the entire village and retell and recount these stories to the village.

11:19

So as scientists, we get so caught up in the science and technology part of what we are doing, you know -- which is the next best model to have, how can we increase the accuracy of my system, how can I build the next best system there is -- that we forget the reason why we are doing this: the people. And any successful technology is the one that keeps the people and the users up front and center. And when they start doing that, we also realize that technology is probably a very small part of this and there are other things in the story. Maybe there are social, cultural and policy interventions that are required, as much as technology.

12:00

So some time back, I worked on a project called VideoKheti that allowed Hindi-speaking farmers in Central India to search for agricultural videos by speaking into a phone-based app. So we went to Madhya Pradesh to collect data for this, and we came back and we were training our models and we discovered we're getting very bad results. This is not working. So we were very confused. Why is this happening? So we looked deeper and deeper into the data and discovered that, yes, we had collected data from what we thought was a very silent, quiet village in the evening. But what we hadn't heard while we were doing this was that there was this constant buzz of night insects, you know? So throughout the recordings, we had this "bzz" of the insects, which was actually distorting our speech.

12:50

The second thing was that when we went there to kind of test our app in the village, I and my colleague Indrani Medhi, who is a very well-regarded design researcher, we found that the women couldn't pronounce the sanskritized words that we had for some of the search terms. So, like ... (speaks Hindi) Which is like the term for chemical pesticides, right? Because we got these terms from the agricultural extension center and the women, even though they are farming, do not interact with that center at all. The men do, the women probably use something much simpler, like ... (speaks Hindi) Which basically means killing pests with medicine. So what I have learned through my journey and what I would like to put across to you -- by now, I hope

you've understood me, is that there is the majority of the world's languages that require intensive investment for resource creation if they are to benefit from language technology. And this is unlikely to happen in a very fast and efficient manner.

14:05

So it is extremely important for us to ensure that the community derives maximum benefit from whatever that we are doing in the language tech area. And to do this and deliver a positive social impact on these communities, we follow what we call the modified 4-D design thinking methodology. So the 4-D means: discover, design, develop and deploy. So discover the problem that language technology can solve for a particular language community. This observation-led approach can help allocate resources where they are most needed, designed for the users and their language, understand the diversity in the linguistic properties and the languages of the world. And don't think, oh, this is made for English. Now, how can we just adapt it for Marathi or for Gondi, right? Develop rapidly and deploy frequently. It's an iterative process that will help you fail fast and early failures will eventually lead to success.

15:07

The important thing is to persevere. Do not give up. And I remember the story of these two Aborigine Australian women, Patricia O'Connor and Ysola Best. In the mid-90s, they went to the University of Queensland and they wanted to learn their own language, called Yugambeh, and they were told very bluntly, "Your language is dead. It's been dead for three decades. You cannot work on this. Find something else to work on." They did not give up. They went to the community, they dug up oral memories, oral traditions, oral literature, and founded the Yugambeh Museum, which became the most important cultural and linguistic center for the language and its community. They did not have technology. They only had their willpower. Now, with the power of technology, we can ensure that the next page is written in Salmi from Finland, Lillooet from Canada or Mundari from India.

16:10

Thank you.