CRAIG: 00:03 Hi, this is Craig Smith with a new podcast about artificial intelligence.

In the biblical book Genesis, humanity is united by a single language, allowing them to collaborate on massive projects, including a tower to reach the heavens. Angered by man’s audacity, God famously confounds their language so they can no longer understand each other and humankind is divided and dispersed among different tongues.

Well, thanks to machine learning, language barriers are falling and man may soon be united linguistically again – this time through simultaneous translation technology. Researchers at Baidu, the massive Chinese tech company, are on the cusp of creating an artificial intelligence system that will allow people from different languages to speak to each other with only a few seconds delay.

Deep learning, as anyone paying attention knows, has revolutionized machine translation. Google’s translation app on your smartphone or the Internet-telecommunications service, Skype, already offer real-time translation of conversations. But those systems translate speech consecutively, sentence by sentence. Baidu has gone further with a system that translates simultaneously from speech to text. Soon, the company says it will offer simultaneous speech-to-speech translation, effectively erasing language barriers.

The technology is in its very early stages, but given the speed at which it has advanced, it holds the promise of uniting mankind linguistically in the not too distant future. I spoke with Liang Huang, the principal scientist at Baidu research in Silicon Valley, about how he cracked the simultaneous translation code and about Baidu’s successful research model generally. I hope you find the conversation as exciting as I did.

I'm here with Huang Liang, or Liang Huang, depending on which direction you want to read it, from Baidu research in Silicon Valley. And you're going to begin by introducing yourself and telling us about your research.

LIANG HUANG: 02:16 Okay, great. My name is Liang Huang and I'm a principal scientist at Baidu research Silicon Valley. And I'm also a professor at Oregon State University. So recently we've made a breakthrough in simultaneous translation. The previous real time translation technologies are real time, but not simultaneous. So what's the difference? Well, real time translation, like Skype is consecutive in the sense that you have to finish a sentence before the translation starts. So it will not start a translation when you are speaking. You complete a sentence, then it calls the translation engine to translate for you. So that's a typical consecutive translation scenario. When presidents of China and the US meet, they employ such kind of consecutive translation using human consecutive interpretation. President Xi Jinping speaks a sentence in Chinese, then somebody translates into English and president Trump responds in English and somebody else translates back to Chinese. So this mode is real time, but it will cost you a doubling of the time.

LIANG HUANG: 03:09 So one hour of conversation now needs to at least two hours to finish. In a conference setting when you use consecutive translation, then it renders the audience largely out of sync with the speaker. For example, the speaker's onto the next slide and you are still reading the translation of the previous slide. So we wanted a technology that is really simultaneous, that is, you're only lagging behind by about three seconds. So a one hour conversation can be done in a one hour plus three seconds. That is the service that many diplomatic situations use. When president Trump and President Xi Jinping met in Florida last year, they actually used simultaneous interpretation, human simultaneous interpretation.

CRAIG: 03:43 And the problem with simultaneous translation is that there's actually a very limited pool globally of people who have that ability to translate simultaneously.

LIANG HUANG: 03:51 That's right. There was a report that there are only 2000 to 3000 simultaneous interpreters worldwide and United Nations has a lot of them, the European Union has a lot of them and each country's department of state has a lot of them. It's extremely difficult for human beings because you have to do the understanding or comprehension in one language and production of speech in the other language. With the limited capacity of human brains, it's just really difficult to do. So each simultaneous interpreter can only sustain for about 10 minutes and there was a study that shows after just minutes of interpreting their errors grow exponentially. The best human interpreters can only retain about 60% of the source language material.

CRAIG: 04:27 So they're actually losing 40% ...

LIANG HUANG: 04:30 They strategically do that for a reason because if they tried to cover everything then they're lagging behind even more. They try to be in sync as much as possible with the speaker so they have to kind of intentionally drop about 40% of the material.

CRAIG: 04:43 Tell me how you got hired by Baidu ...

LIANG HUANG: 04:45 So I got hired by the senior VP of Baidu, Dr. Haifeng Wang, who was a very senior person in natural language processing. We have known each other for more than 10 years. So he approached me and hired me. And the first task he suggested I work on is the simultaneous translation project. And I knew this problem was ridiculously hard and kind of the holy grail of artificial intelligence. And I actually worked on it for about half a year at Oregon State University before I joined Baidu. I knew it was extremely hard. So I cannot promise that I can solve this problem, I can deliver it, but I, I will spend my time to, to think about it. And then I got the Eureka moment.

LIANG HUANG: 05:20 So this is so difficult mostly because of the word order difference between languages. So think about German. Maybe many of you understand German and you know German has a very peculiar word order, which is called verb final. In the embedded clauses, the verb comes at the end in the matrix or main clauses, the verb comes at the second positions. But English is SVO, which is subject, verb, object. Very normal. Chinese is a mix of SVO and SOV. So when you translate from German to English, you often have to wait a long time, about five seconds or 10 seconds until you've heard the German verb.

LIANG HUANG: 05:52 There is a famous joke in the European Union or maybe even the United Nations that whenever a German speaker speaks, the interpreter is often got stuck and doing nothing. And people would ask the interpreter why you're not translating? And he'll say, I'm waiting for the German verb! Chinese is kind of like that. In Chinese, the verb often comes at the very end. Chinese is a mix of SVO and SOV.

LIANG HUANG: 06:13 In this particular sentence, 'Bush President in Moscow with Russian President Putin meet,' the verb 'meet' comes at the very end. But in English we have to translate it into President Bush meets with Russian President Putin in Moscow. So if we don't want to wait until the German verb or until we've heard the Chinese verb, what can we do? So we have to be more in sync with the speaker. We cannot be lagging behind him for a long time. So what I realized suddenly one day when discussing with my student, MingBo, who is the first author of this project, is that sometimes we can be anticipative.

LIANG HUANG: 06:43 We can go ahead of the speaker and predict what they're going to say in about five seconds into the future. And we can do that with pretty good confidence. Because once you heard 'Bush President in Moscow' for example, what can he do in Moscow? He has to meet with somebody, likely Putin. With enough information, with enough training data, this kind of context, not just President Bush, but maybe President Obama or President Clinton, and not just Moscow but maybe in Paris or Beijing, meet with somebody. So you have a lot of sentences like this in the training data. You know that the president of the United States in Moscow is likely to be meeting with somebody so you can predict a verb, meet with pretty good confidence. So this is my Eureka moment. I got this kind of wait-K model. It's called a prefix to prefix model, where you don't rely on a full sentence, you just rely on the first say two words.

LIANG HUANG: 07:27 First two Chinese words to predict the first one English word and the first three Chinese words to predict the first two English words. And when you have the first four Chinese words, 'Bush President in Moscow,' then you go ahead and predict the third English word is meet - long before you heard the Chinese word for meet. So this is anticipative and by doing so we were the first system that can achieve latency requirements. For example, if I want you to be just within three seconds delay or would they say four seconds delay, now we can do that because of this wait-K model with anticipation.

CRAIG: 07:56 And so you set the requirement ...

LIANG HUANG: 07:58 The user can set the requirement ...

CRAIG: 08:01 ... you set the latency and then the loss factor depends on the latency. The loss factor goes up if the latency is ...

LIANG HUANG: 08:07 There's a trade-off between latency and quality. With more latency, you have better quality. With low latency you have expectedly lower quality. It also depends upon language pair or the language direction that you translating. For example, if you're translating between French and Spanish or between Spanish and Italian, you probably only need to wait one word or you can do very low latency and almost word by word translation. But if you're traveling in between German and English, especially from German to English, you have to wait longer. Say from Japanese to Chinese or from say Japanese to English, you also have to wait longer. Right. It really depends on the direction. From English to German you have to wait less than from German to English.

CRAIG: 08:42 And what happens, for example, if the verb turns out to be wrong, does it self-correct?

LIANG HUANG: 08:48 That's a very good question. So human interpreters, when realizing a prediction error, they attempt to correct. If it's a major error, if it's roughly okay, like a paraphrase, they would not correct it because attempting to correct mistakes in interpretation would cause them to be more lagging behind, more out of sync with the speaker. First of all, they only do to anticipation when they're quite certain. And in the cases where they made a mistake, they attempt to correct it. So far our system does not have the capacity to correct mistakes. We don't correct mistakes like this, but in the next version of our system we will. So we are actively developing ways to detect and correct mistakes.

CRAIG: 09:21 And will that, as in dictation, then erases and rewrites.

LIANG HUANG: 09:25 So there's two options. One is it erases and rewrite that would work for speech to text translation because the text display you can always rewrite. We currently don't do that for two reasons. One, constantly rewriting the translation would be kind of annoying to the user because the user sees constantly going forward and going backward. Secondly, more importantly, we are anticipating a more important application, which is speech to speech translation in the future, maybe next year. And in speech to speech you cannot retract a word which you have already said. This is sound which you cannot retract.

LIANG HUANG: 09:56 So we are kind of thinking in that direction where we can't actually erase and retranslate. We had better just to say, oh I'm sorry I made a mistake or whatever. We're developing those kinds of technologies.

CRAIG: 10:05 And talk about speech to speech. You're also working on that in Silicon Valley?

LIANG HUANG: 10:09 Yes, but that's a more long-term project.

CRAIG: 10:12 What are the challenges there? From the science point of view.

LIANG HUANG: 10:15 It's a major challenge that we have not been able to do yet is the missing step from text to speech. That is from the translated text to the translated speech. So currently there are three steps in speech to speech translation: there is speech recognition, which recognizes source speech to source texts. So from Chinese speech to Chinese text, Chinese characters. What we do here in particular was translating from Chinese text to English text in a simultaneous way.

LIANG HUANG: 10:40 But there is a third step that is generating speech simultaneously from English text so that you can complete the loop from speech in Chinese to speech in English. So the third step, simultaneous text to speech synthesis is the thing that is missing. And that is also very challenging and we don't have a good way to do it yet, but we are actively working on it.

CRAIG: 10:58 Although the latency from text to speech should be less ...

LIANG HUANG: 11:03 Should be less, but still current technology also relies on the whole sentence, right? Current text to speech synthesis relies on the whole sentence to do it because you have a full sentence, then you can do the parsing very well. Say for example, you know at the end of the sentence that it's actually a question sentence, then you do rise it, right? But if you don't know it at the beginning portion of the sentence, then you cannot plan very well.

LIANG HUANG: 11:22 You have to have some knowledge of the global structure of the sentence or some prediction of the future.

CRAIG: 11:23 This predictive model that you have, what is the error rate?

LIANG HUANG: 11:31 In Chinese to English translation, at least we have some numbers. A wait-three-word model, roughly speaking, that will translate to waiting about 1.5 to two seconds - so about two seconds behind these Chinese speech - would have some, some degradation in quality. But if you wait five words, the degradation is less. So five words is, roughly speaking, three seconds because Chinese speech is roughly two to three syllables per second, and Chinese word is always about two syllables. So roughly speaking, the take home message is that our simultaneous translation with about three seconds delay is about 3.4 BLEU score less than full sentence translation - full sentence translation, meaning consecutive that you have the full sentence, then of course you can do a better job.

LIANG HUANG: 12:12 So we're about 3.4 BLEU scores less. Um, but we're constantly improving that number.

CRAIG: 12:18 What is [BLEU score](https://en.wikipedia.org/wiki/BLEU)?

LIANG HUANG: 12:19 BLEU score [Bilingual Evaluation Understudy Score] is the standard metric of translation quality. By comparing machine translation output with the human translation reference. So for example, I have 10 sentences in Chinese and then I ask a machine to translate into English. I got 10 English sentences output. And then I also ask human beings to do these translations and I compare their translations. How close is your machine translation with, you know, to the, to the human translation, if it's very close, you use a lot of same words, the same phrases, then you got a higher BLEU score. If you are very different, they had no word is the same, then you got zero BLEU score for example. So this 3.4 BLEU score is a noticeable difference. It's definitely noticeable difference, but it's the price you have to pay if you want to achieve simultaneity. Right. So there's again a trade-off between latency and quality.

CRAIG: 13:02 But you would call it acceptable.

LIANG HUANG: Acceptable, yeah. Acceptable kind of degradation in quality than the full sentence translation.

CRAIG: And given that you've only been working on this since ...

LIANG HUANG: ... since like April.

CRAIG: You would expect that that would improve.

LIANG HUANG: 13:16 That would improve, yes. Yes. And this is really the first system, the first approach that can achieve a given latency metric. Previously you cannot say, oh, please be at most two seconds delay or please be at most four seconds. You cannot specify that. No. No previous approach can achieve a given latency requirement.

CRAIG: 13:33 Yeah. How long do you think before this is commercialized?

LIANG HUANG: 13:38 It's kind of our internal product by now and we use it in our Baidu world conference, which is a public conference. A lot of international conferences are happening and a lot of them want to have simultaneous interpretation or simultaneous translation services, but it's extremely expensive. Like as I said, the availability of such qualified interpreters is, is a problem. There's a huge demand for it, right? There's not enough supply

CRAIG: 13:57 and, and you think speech to speech is a year way or ...

LIANG HUANG: 14:02 At least half a year away. At least half a year away.

CRAIG: 14:03 That's amazing.

LIANG HUANG: 14:05 It's, it's ridiculously difficult. Even more difficult than the simultaneous text to text translation we're doing right now. So five words or about three seconds delay is still a lot better than 12 seconds delay or one sentence delay in the previous technology.

CRAIG: 14:24 I wanted to talk a little bit about Baidu research in general. How much research in China today is being seen? There's much less focus on publishing and a lot more focus on getting products done. And so the question is, are there advances in China that we don't see?

LIANG HUANG: 14:45 That's a great question. I think there's a lot of advances in basic research and in applied research in China. Not just products. Of course we made great products like WeChat, like Alibaba, Baidu search, a lot of stuff. But China also invests in basic research. So you can see that a lot of uh, not just Chinese authored papers, but also only mainland Chinese authored papers from universities or companies within the mainland China is increasing very rapidly. In the past, maybe 15 years ago it's a very, very low percentage, but now it's huge.

CRAIG: 15:13 That's right. That's all open research that everyone sees. Is there research that is not being published because companies are just too busy to publish.

LIANG HUANG: 15:23 That's also possible. That's also possible. But in Baidu, we still encourage publishing. Another reason why we have an office in the Silicon Valley doing a little bit more long-term research, this kind of geographic and time zone separation between Beijing and Sunnyvale is huge. Sometimes it's a huge disadvantage in collaboration, but sometimes it's, it is an advantage that we have slightly more freedom in thinking more long term. People in Baidu research in Sunnyvale focus more on long term directions on hard problems, like the hardest problems that nobody knows what the answer about.

CRAIG: 15:53 But, but generally this question of visibility into what's happening in AI in China. You think if someone like me follows archive, I see pretty much everything that's happening in China?

LIANG HUANG: 16:04 Yeah, a lot of Chinese companies and the universities [arxiv](https://arxiv.org/) their papers, like this paper was arxived in November. So Chinese companies and universities definitely do a lot of publishing for sure.

LIANG HUANG: I myself don't think there is a reason to hold on to publish. Maybe some companies do, but Baidu doesn't. Many other projects in Baidu research, even got open sourced, and so our researchers released their code almost immediately after the paper was published or even after immediately after paper was arxived.

CRAIG: 16:29 This is open source?

LIANG HUANG: This is not open source yet, but the idea is so simple that anybody in the world can implement it very quickly. Anybody reading the paper will be impressed how simple that is and seemingly it is, but actually it's very deep. But, but the implementation is quite easy. So, in a sense, it's very close to being open source.

CRAIG: 16:45 There's been a lot of talk about this. China's pouring a tremendous amount of money into AI research, the US, at least in the private sector, is pouring a lot of money into AI research. Who's ahead or who's going to be ahead? How do you feel about that? You straddle both worlds.

LIANG HUANG: Right, so I think Baidu is very open to collaboration with not just companies and universities in China, but also companies and universities in the States and outside of China in general. So we are very open to publication to collaboration with universities in the States and Europe.

CRAIG: 17:15 I'm trying to figure out just personally or as a journalist, whether that's a red herring, whether there really is competition.

LIANG HUANG: I don't think there's a competition in a sense that research is all about building on somebody else's work, standing shoulders of the giants. Either you are in China or in the US, you read papers from both sides. Each of us was just building up on other people's work.

LIANG HUANG: 17:36 To me there's not such competition in AI. And I think it's a good thing that both governments are investing a lot and both private sectors in both countries are investing a whole lot in AI. That's great for both countries. The technologies developed here can be used there and vice versa. It's all open source, all published.

CRAIG: Where do you think this simultaneous translation will go? What's your vision?

LIANG HUANG: 17:55 So the next step, as I said, would be simultaneous speech to speech all the way. So, right now it's only speech to text. If you can do simultaneous speech to speech, that's much, much better. Baidu has a device called Wifi translator, but it's doing consecutive translation right now. If this device can do simultaneous translation, then you can save a lot of time by traveling to Japan or Germany. Right now you use this to order food in a restaurant, but it costs you twice as much time.

LIANG HUANG: But now with simultaneous translation, say if you go to a conference employing our technology and the conferences in German, then you can still understand it very well. With this consecutive technology, you cannot date someone. With simultaneous interpretation, you might be able to. If you have a headset and the left track is speaking Chinese and the right track is speaking English, then you can do this simultaneous interpretation. You can talk to another person in different language completely not understanding each other's language and can talk to each other and probably even date each other. And that will be wonderful.

CRAIG: That's fascinating, yeah.

LIANG HUANG: 18:47 Even though this is kind of trying to replace human beings, it's not our intention. Human interpretation is still needed in many situations. Like maybe President Xi Jinping and President Trump meet because they need very high quality translation and the machines are not there yet. We are trying to make this technology more available and to help reduce the burden of human interpreters, not to replace them.

CRAIG: 19:09 For those of you who want to go into greater depth about the things we talked about today, you can find a transcript of this show in the program notes along with a link to our Eye on AI newsletters. Let us know whether you find the podcast interesting or useful and whether you have any suggestions about how we can improve.

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