Dr Wan Mingyu - Corpus Linguistics Meets AI: The Next Language Revolution

Welcome to this episode of Exploring the Humanities, Voices from the Hong Kong Polytechnic University. Our podcasts allow us to showcase the exciting and innovative work being done by our colleagues in the humanities. Intersecting with fields as varied as aviation, mental and physical health, virtual reality, religion, GenAI, design, neuroscience.

I am Renia Lopez from the Faculty of Humanities and with us today we are very privileged to have Dr. Wan, Clara, lecturer in the Department of Language Science and Technology of The Hong Kong Polytechnic University. Clara, thank you so much for being with us today. To start with, tell us a little bit about your background.

- Thanks for inviting me. My background is multidisciplinary. It sits at the intersections of corpus linguistics, computational linguistics, natural language processing, and importantly, I have an interdisciplinary academic background training in electronics engineering. That gives me a quite good foundation of training in logical thinking and systematic viewing of certain research questions.

Could you give us a broad overview of your research please?

- My research began with corpus linguistics, which aims to study language patterns through the use of corpora, that is a collection of textual data. Through corpus linguistics studies, we use techniques such as concordance, keywords in context and queries. This can help language users or teachers find prominent language patterns. This provided the foundation for my current research on computational linguistics, which is based on the modeling of linguistic phenomena via various kinds of computational models where I focus more on the natural language processing component. As we know, computational linguistics is quite a broad area that can cover a lot of issues and techniques. My research focuses on modeling the intrinsic and implicit language footprints in all forms, in combination with the existing state of art models like embedding and transformers, which enable the current intelligence of large language models.

You mentioned that you work with corpus linguistics.

- Yes.

What is that?

- To make it simpler, it is collections of data in textual form, most of the time, and using techniques of concordance to find the collocations of language patterns. It is usually used by people such as language teachers to try to find the significant collocations with a word like "go with", "go viral with" in order to teach and also to showcase to students how language works. It is a technique to show keywords in context, to provide frequency information for the students to know the patterns of collocations.

This is really useful for a second language learner, so someone who is trying to learn English, for example, and to work out all the possible uses that you can have of the word "go", as you've mentioned.

- It is useful for almost everything you want to learn about or if you're going to write an essay and you are not sure how to write the words properly. In these situations, collocation works.
- But this is just the earlier research of mine, done while I was a PhD candidate. Now my work focuses a little bit more on computational linguistics. It is actually based on corpus linguistics, while it is also multidisciplinary, incorporating and combining more techniques, such as machine learning, natural language processing, NLP in short. NLP actually is the core component of current language models. So, you have probably tried to use ChatGPT, right? These models are actually based on training the intelligence of machines to be able to understand language, the meaning, to process the language and then generate respective responses to the inquiries of the users. This actually summarizes the core components and techniques for enabling machine intelligence.

Are you saying that corpus linguistics is basically the basis of all of this computational linguistics and NLP work that people are doing today and by extension AI and the models, the language learning models?

- You could say so. It's like the earlier foundations of this trajectory leading to the current powerful intelligence of the models, from language models to large language models to the current advances of transformers. They're actually dealing with the core common core issue of understanding and representing language.

Thank you. I think that made it much clearer for our audience. We often get students asking, what exactly is corpus linguistics and how does that work with AI? So, it's good to have this basic

knowledge. You've mentioned computational linguistics. Tell us a little bit more about computational linguistics. What does it actually mean? How many different subfields are there within computational linguistics, and which do you look into?

- It actually encompasses a wide range of topics and different methodologies. Generally, it uses computational techniques, trying to study language and understanding it. But this is a broad issue. My research focuses more on modeling linguistic features, the footprint where the interface of language forms, like the patterns seen in the frequency of words used.
- These features are the salient patterns that can tell us how the language behaves and show the traits of humanistic behavior, communication patterns and even societal behavior. In short, my research focuses on enabling this model by incorporating implicit external knowledge on the basis of accessing this textual data, in order to enhance the model's sensitivity to not just word and meaning but also to understanding the real emotions, cognitions, implicit knowledge like figurative language metaphor and the lies, sarcasm, all these interesting humanistic things.

You've mentioned metaphors, lies, figurative language...all of those aspects are related to emotions and what is, in effect, affective computing. So, you're looking at the affective side of the language. Is that correct?

- That's one of the focuses of my research. But if you're asking about this computational linguistic research scope in a broader range, it is not limited to just affective computing. I cover it because it attracts me and drives me to really conduct some research, trying to find out the intrinsic mechanism in machine intelligence in understanding language.

Do you think that's possible? Do you think that a machine can actually get to the point of understanding these nuanced meanings that we can have human to human?

- It is happening. It is becoming a reality now. It's not easy—it's a challenging task that involves a lot of work, such as solving the difficult problems of context modeling. We know language is dynamic and it's complex. So that's also part of the reason that attracts me to conduct such research. And a group of us in the community of computational linguistics have done successful work to show that it is possible.

Let's get into the ethical side of things for a minute. Do you think we want to do that? Do we want machines to behave like humans do?

- That's a very important issue actually. This is a common problem that actually everyone in the

world is encountering. Machines can be incredibly intelligent and powerful given the current techniques, such as: incredibly big-data inputs into the model training with highly advanced machine learning techniques, Deep Learning, world knowledge representational techniques with embeddings, external and cognitive data resources... People are doing all of these works together. That's leading to what you're witnessing and seeing and using now: the ChatGPT models and transformers. We can't say that this work is trivial. It is happening and then is advancing all the way around.

Are we worried about it?

- I tend to be positive. There are two different voices. Overall, some people are concerned that AI can be too powerful. And it could be so powerful that it could teach itself to evolve and ignore the humanistic needs. In the end, it will probably go in the opposite direction by destroying humanity, once it is powerful enough, like some of the scenarios you see in fiction. That's possible. But as a linguist, a computational linguist in the Faculty of Humanities, I'm very confident that if we address these ethical modeling concerns with proper regulations and infuse this into the modeling of AI, which is actually essential and necessary, we will guide the AI to be a more humanistic intelligent system with mercy.

You had mentioned before lies and metaphors. How does a computer being able to process lies and metaphors help it to be a better system?

- I would like to say that machines are not able to understand language. Allow me to talk about more technical issues: How current agent systems are able to understand a question is largely based on the representation of natural language into vectors and embeddings. So, it's basically numbers and all dimensions of numbers based on statistics to find the probability of what exactly people are saying. And that is literally. It is limited to formal representations, largely based on speech, texts or current multimodal representations. It is powerful already, but still, it is based on the surface form of what we can input to the machine in terms of sounds and texts and images and videos.

So, you're saying that every single word, every single context, every single image, every single sound is actually transformed into a number?

- Yes. This also shows the reliance on such-representations in order to empower or enhance the current model's performance. So to go back to your question, why would I be interested in

studying those less implicit items of information like emotion, lies, sarcasm and metaphor? Metaphors are figurative and more pertaining to humans' cognitive thinking about language without access to any surface information. One of my papers [on metaphors] mentioned the sensory-motor [aspect of metaphors]. The ideas are based on cognitive theory and cognitive linguistics, trying to understand language as an interface of people in their mind connecting to the external world via the interfaces of the six senses: Vision, audio, touch, taste, sound, etc., and actions like body movement, hands and feet. This information changes all the time. And it's hard to feed this information holistically to the models to make them able to perceive the external world, the space information all the time.

Because what you're saying is that every single one of these inputs has to be transformed into a number?

- The current models are based on numbers, because our computers only works with zero and one.

Then what happens if you have managed –looking at this room around us–, if everything in this room has been coded and every possible action or interaction that we have with the objects and with each other in this room has been coded? Is that enough then for the machine to be able to interpret more abstract meanings like that which a metaphor might have?

- That requires our collaboration as human beings together. That's why in the humanities, in the Faculty of Humanities, people work in different sub-disciplines, some focusing on more neuroscience, where they rely on setting up behavioral experiments like trying to get connected to the human mind with access to EEG or to catch eye movement to understand human behavior while they are producing language. That's one group of people, and there are people focusing on testing different dimensions of such implicit information with the existing representations, like do a fine-tuning. What we can do is limited, but still, we're on the way to find more different directions of understanding language truly and intrinsically.

In this paper you've mentioned, where you have been studying metaphors, tell us more about how did you do it and what aspects of the sensors or actions that a human can do did you take into account?

- As I mentioned, metaphor is also a kind of implicit figurative language. The idea in my paper is that I try to utilize an innovative methodology to study metaphor. I know most of the existing

researchers, even now, are still heavily relying on manual annotation and analysis to study the domain mapping phenomenon, different sources and different categories of sources of metaphors.

Tell us a little bit more about the mapping and the different sources and give us a couple of examples for our audience.

- For instance, I would use examples more related to sensory input because I use sensory diction and norms. For instance, “sweet” voice. This is a very simple, adjective-noun phrase. This “sweet” is used to depict the sense of taste but voice is a noun referring to the sense of listening. This domain transfer from using the taste sensory dimension to the audio one is a very good example of a metaphor. Metaphor is actually everywhere in our conversation. Even if you're not an academic, you are actually using metaphors all the time. That's the early reason driving me to study metaphor. But then mine is not the traditional way of studying metaphors—I'm not saying they're not good—, because I use computing methods. I'm trying to automatically detect metaphors in texts and then utilize the existing data sets to train the models, trying to understand the metaphors’ distributions across different groups. Those were my early objectives. When you start doing research there are a lot of circumstances that leads one to explore other possibilities incorporating external knowledge. And sensory motor is the one I tried in my research.

So why do we use metaphors? Why is it that you will see people using them more often in specific contexts?

- Yeah, that's a very ambitious problem theoretically, I'm not a theoretical linguist, but it's interesting and a thought-provoking topic I have noted while doing this research.
- I think it's natural when people say metaphorical things. In my understanding, people are trying to express information more efficiently for more effective communication. By using metaphorical expressions they are able to access commonly shared information between the speakers by accessing more reachable source domains. I didn't aim to contribute to the theory, what I tried to do is to make the model work more like a human. This is a big step to reach, even though I have only done a little bit, I think that's already satisfactory. I'm happy with that. Of course, we have a lot of other issues to solve in order to answer such questions. I believe we don't have a final answer or a universal answer, but I think metaphor itself is intriguing. It can help us understand natural language better if the models can be trained with such datasets and theories.

You had also talked about lies. Tell us more about the work that you're doing with lies. Are you also trying to understand why we lie?

- The original purpose is not trying to understand these theoretical questions as I said. My work is based on my observation of the infodemic.

What do you mean by infodemic? What is that about?

- Infodemic is a linguistic term, a combination of information and pandemic. This blended word was quite innovative at the time, indicating the rapid transmission of lies, misinformation, rumors and all kinds of misleading information online during COVID-19. COVID-19 itself was a pandemic. So it's like a metaphorical expression of how such information, such lies and misinformation, spread like a pandemic.

Two interesting topics here. One is this new blended word that came about during the COVID pandemic - or was it around before that, "infodemics"?

- The term "infodemic" is kind of innovative. Quite a handful of researchers did intensive work during the pandemic and post-pandemic on it. But the phenomenon of infodemic exists since human history started. As long as there are humans, there are lies and people who like to gossip. People tend to spread lies more than they spread positive information. And that's one of the interests of my paper, the paper about the infodemic. While I was doing this preliminary research, I found that there is a coordinated effect between emotion and infodemic, this fake news spreading and dissemination. As my paper suggests, negative emotions such as fear and anxiety tend to drive people in disseminating this information, which makes the infodemic more severe.

You had mentioned that lies spread faster than facts. You have also mentioned that it's the negative emotions that seem to drive the spread of this negative information. Why are we talking about negative emotions and why are there no positive emotions spreading information?

- Actually, my studies covered all kinds of information, including negative and positive parts. With data-driven and statistical findings, my study supports the observation that negative emotions such as fear and anxiety tend to drive people to disseminate and spread fake news more widely, faster and more often. This is also supported and accounted for by my later theory in psychology, a prospect theory, trying to use the gain and loss framework to account for such a phenomenon. Because intuitively, if you see something negative online, especially if it is health related with

high stakes where people might be in the situation of potential life loss, people tend to believe the information and then spread it quickly. And that causes the infodemic. So it is related to people's actions, a societal action encountering risks, gain or loss. So if there is the risk of a loss, people tend to disseminate it. I have a framework theory on that in my paper.

Are we as humans more worried about what we might lose rather than about what we might gain? So if we are in a situation where we might gain a lot, but lose a little bit, we would prefer to lose a little bit rather than the chance of gaining a lot. Is that the gain and loss theory?

- I think it makes sense. Yes, but I wish I had enough data to support that. This is based on the statistical analysis and all the modeling things we found: That negative emotions are prominent in driving such infodemic in language patterns. We need more theory to try to understand this behavior, why people tend to take actions while they fear a loss instead of a gain. People might feel less motivated to share the gains when they feel they have gains. But then when it is a risk of a loss, they want other people to share the loss, if I'm right. But I'm not an expert in psychology.

Do you think that we want to share our loss or that we, in a bizarre way, enjoy other people's loss?

- Kind of. Do you think so?

I do actually believe that we have a nasty streak to us.

- Gossiping. To share the pain. Yes. I didn't mean that people don't share the positive things, but usually they get paid to do that (like people working in departments in government).

You certainly don't hear a lot of positive news when you switch on the news, for example,

- Think about it: The data is based on social media data people create in a free mode. They are not paid to do this. And that shows the natural way society behaves. That's why I like to study social media data in my research.

And I suppose all of our social media is the corpus that you use in these corpus linguistics, computational linguistics exercises that you carry out.

- Yes, quite a big group of us are doing this kind of research using social media data as a corpus. It's called Web-as-corpus, WAC in short. This is one of the types of researchers working in linguistics, in corpus linguistics. Actually, we cannot set apart the fields strictly. There are certain overlapping areas between corpus linguistics and computational linguistics and even AI.

One last question for you. In many of your papers, you talk about sentiment analysis. I imagine it's related to some of the things that you've already told us about, but what exactly is sentiment analysis and how does it work within the research that you carry out?

- Sentiment analysis is closely related to emotion and affective computing, they are in a hierarchical relation. Sentiment is more focused on studying the polarity of language. While you're saying something, it tends to show positive or negative or neutral polarity in your expression. Emotion is related to sentiment, but it has more dimensions of emotional aspects of language patterns, more fine-grained categories instead of just negative or positive polarity. It could be like what I studied in my research: Multiple categories, emotions, sadness, disgust and so on and so forth. And in theory, it also incorporates multiple features, including violence, affection, empathy and sympathy. And all of this together can contribute to the research of affective computing. To answer to your question about sentiment analysis, we are working to understand the affection, the sympathy, emotion and empathy of human beings through the modeling of such language patterns.

You have talked about many different areas within linguistics, applied linguistics more generally. I think for any future students of the humanities, you've given them many ideas and many future research projects to get their hands into. Are there any last words that you would like to tell our audience?

- Yes. Thank you, Professor Lopez. I wanted to summarize all of these research topics I have been studying. It looks like they are quite diverse but actually, my research serves one ultimate goal: That is, trying to understand language from a humanistic perspective, instead of just the superficial representations of language to mimic the human language behavior. My research tries to really understand the soul of language, the mechanism of language. My research is not based on separate pieces of research, it is systematic. I have a passion that drives me to strive hard to work in that way to find proper solutions and answers to account for complex issues in language. My teaching is also aligned with my research goals. This will also help students to understand more easily and profoundly these multidisciplinary areas of language-related subjects, including linguistics, corpus linguistics, applied linguistics, language teaching and learning, natural language processing, even computer science, artificial intelligence and generative AI. I find that all these areas help to answer the same question - what is language? What is humanity?

Thank you very much, Clara. And I think for our listeners, if you thought that the humanities were dead, this has probably shown the opposite. Not only that, I think there is a lot of future for the humanities based on all of these areas that you've told us about. We encourage any young people out there to come and explore these topics. Thank you so much, Clara, for joining us today and for sharing all of your work with our listeners. It has been a pleasure to have you on this podcast.

It's been my pleasure.

Thank you for joining us on exploring the humanities from the Hong Kong Polytechnic University. For more episodes or information, do visit our website or follow us on Spotify. Stay tuned for discussions with leading voices from the Faculty of Humanities and beyond.