INTERVIEW WITH DR. RUIHUA SONG

A photo with part of my team members on Dr. Ruihua SONG's birthday in 2022 (the fourth person in the left is Dr. Ruihua SONG)

Dr. Ruihua SONG is now a tenured Associate Professor of Gaoling School of Artificial Intelligence, Renmin University of China. She worked at Microsoft Research Asia from 2003 to 2017, where her paper on web page block importance modeling got Best Paper Runner-up Award from WWW 2004 and her proposed generic main article extraction across websites algorithm was shipped as reading view features of IE 11.

On May 2017, the first poem collection that is 100% created by an AI was published (ISBN: 978-7-5596-0296-1). The book title is "Sunshine Lost Windows". Her team contributes the generation algorithm from an image to a poetry. Later she worked at Microsoft Xiaoice team as a Chief Scientist for three years.

In September 2020, she joined Gaoling school of Artificial Intelligence, Renmin University of China. As a principal investigator, she works on WenLan project and have delivered a series of large-scale Chinese multimodal pre-training WenLen models. Her current research interests include multi-modal understanding, interaction, and generation. She served SIGIR and EMNLP as Senior Area Chair or Area Chair and also Information Retrieval Journal as Chief Editor.

The first poem collection created by an AI and its traditional Chinese version

What are your team's main objectives and responsibility-ies?

My dream is to create an alive AI. By ``alive'', I mean that the AI can grow by absorbing a huge amount of updated data from the Web by machine learning. For example, WenLan models can learn from 650 million image-text pairs crawled from Internet and map semantics cross images and text into the same space. Now we can search a poem or a song by any input image no matter whether the poem or song has an image surrounded. By ``alive'', I mean that the AI can interact with people face-to-face, which means that we can see it (including its expression, actions, etc.) and at the same time it can also see us and our surrounded environment. For example, WenLen 3.0 video-text model enables a Cyberdog and an AI called Mali to react when seeing real visual scene or give appropriate comment with vivid expressions and actions. By ``alive'', I mean the AI can imagine and create something truly novel. For example, we have done a work to simulate the creation procedure of a human writer and use a series of image and text experiences to guide a long poem generation. To achieve this goal, my team members are encouraged to have experiences of at least two modalities among natural language processing, vision, and speech.

CyberDog is scared by a spider. Its reaction is predicted based on WenLan models

In your opinions, what are the opportunities and challenges for your research topic?

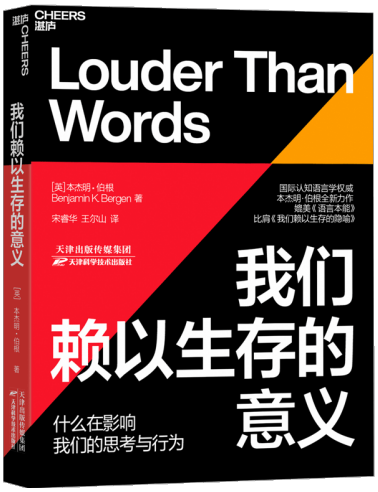
The fast development of self-supervision learning and the success of using Transformer in vision, speech, and language give us great opportunities to multi-modal understanding, interaction, and generation. For example, by using contrastive learning, WenLan 3.0 model can map both a video and comments into the same space and thus the problem of multi-modal dialogues can be converted into the classic problem of dialogue based on question-answer pairs. This opens new opportunities to implement a more natural multimodal interaction system.

As multi-modal dialogues are not necessarily related to vision, there are two main challenges: one is how to get large-scale data that are not intentionally created by crowdsourcing but naturally generated by Internet users. The other is how to bridge the gap between textual dialogue model and cross vision and language model while keeping both merits to respond with appropriate emphasis.

For the above challenges, how do you deal with them?

For the data challenge, we find short videos and user comments are good sources to simulate multi-modal dialogues if with careful cleaning. We are collecting such a dataset and plan to

share with the communities. For the challenge of leveraging both textual dialogue model and vision-language model, we propose some methods to project the unimodal space into the multimodal space and get some promising results for a multimodal text generation task. We are also investigating how many turns among a dialogue are related to vision on naturally generated data, when the multimodal dialogue model or the textual dialogue model plays more important role, and how to combine them to achieve better effectiveness.



The Chinese version of Louder than Words

Currently, big model (wenlan, CLIP, DALLE) has a huge impact on research community, and changes the paradigm for research. From your perspective, what will AI and deep learning be in the next 10-20 years?

Self-supervision paradigm breaks the limitation of manually labelled data. Thus models can consume big data from the Internet. Thus it learns more about the real world from naturally generated data. Such big models have been used in real applications and have more potentials. From my perspective, there are three important things are going to happen in the next 10-20 years: 1) AI can naturally interact with human beings face-to-face as multi-modal information are going to be better understood and generated; 2) AI can walk through the real physical world when it can understand the world in an embodied way; 3) AI can create something truly novel, perhaps a story or a short film.

Do you have any encouragement words for young scientists and students?

In addition to being a good builder, try to become a good dreamer. Artificial Intelligence research needs great imagination to leap again.