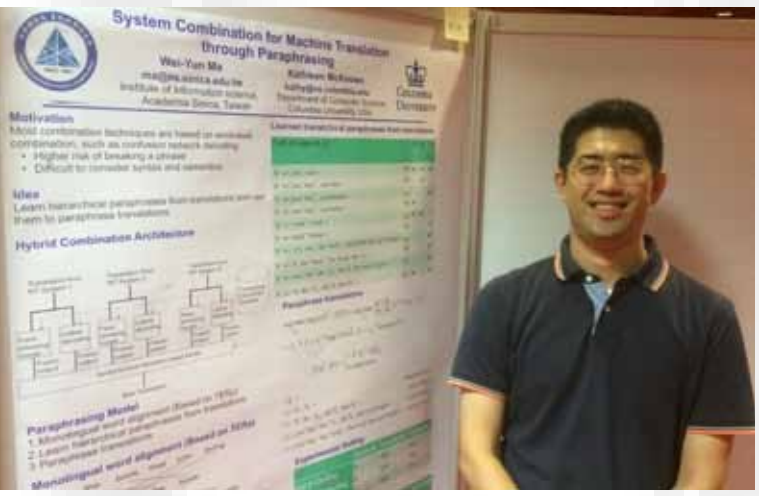
Natural Language Processing – New Developer for Next Generation

Dr. Wei-Yun Ma, Assistant Research Fellow



Presentation of my research in EMNLP conference.

Why did you choose the IIS and what led you to do research in this field?

I first joined the Institute of Information Science (IIS) in 2001 as a research assistant working on natural language processing (NLP). Then I went to the United States to pursue a doctorate and came back to this big family of IIS as an Assistant Researcher. When I was still a research assistant, I felt the enthusiasm and passion of doing research through listening to lectures and academic communication with scholars and researchers. It was from that time the seed of pursuing research as my lifelong career was planted in my heart. The topic of my Ph.D. study was NLP. Upon graduation, I happened to learn there was an opening in IIS. It was a once-in-a-lifetime chance for me. If I could host a lab, I would have a chance to realize many innovative ideas and thoughts. It will be very exciting! IIS is a renowned research institute. It provides a great environment for research in many ways, including (1) the freedom to choose research topics, (2) the presence of at least eight research teams of more than forty researchers provides a great chance for cooperation among different fields, (3) administrative support/grants for researchers, (4) the opportunity to network with domestic and international professors and researchers, and (5) no teaching obligations, so that researchers can concentrate on research. However, its TIGP programs give researchers a chance to teach.

Self portrait

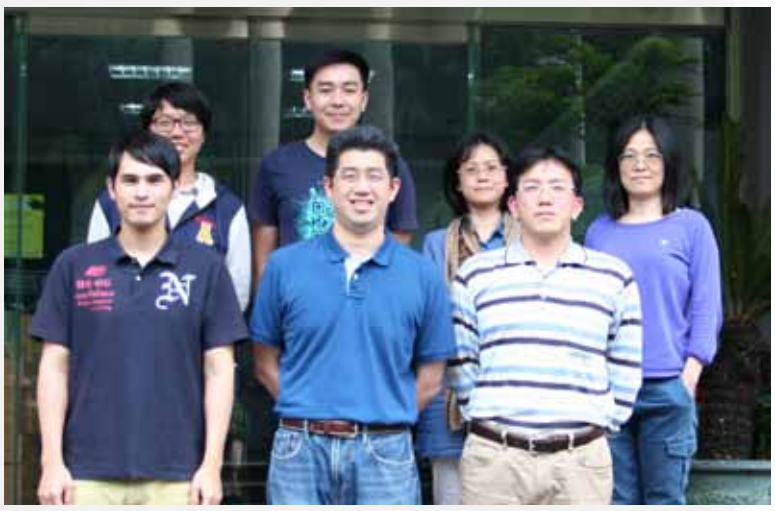
I started taking programming courses when I was a freshman in college. In my junior year, I worked as an intern at ITRI to design a voice command recognition system for the 8051 chip. It was the beginning of my research interest. Later, when I continued the study of voice recognition, I needed to use language models and was surprised to discover that NLP is a field for computer science! It was an eye-opening experience for me. Then I served Defense Industry Reserve Duty at CKIP in IIS for four years after I graduated from the master's program at NCTU. The many NLP studies I had conducted triggered my desire to further my professional skills abroad. Therefore, after my service was completed, I went to Columbia University in New York City, where I worked under Professor Kathleen Mackeown. As I continued to study NLP, I not only gained knowledge from many professors and world-renowned scholars but also learned from them about ways of doing research. While at Columbia, I was fortunate to have joined many intercollegiate grants, such as a multilingual QA program hosted by DARPA and machine translation by NSF. I had chances to participate in big research programs and to learn how they were planned and carried out. During that time, I had many chances to communicate and associate with other professionals from other schools. Before I finished my doctorate, I served as an intern at Microsoft, where I witnessed how a big company valued the importance of NLP research and how related skills have been put to use in services and products.

What are your main research topics and objectives?

Human beings use languages to record knowledge and communicate with each other. NLP is a technology to give computers the ability to handle human languages. Human language is very complicated for computers to understand, in part because there are many ambiguities. For example, when in Chinese we say "Wǒ kǎoshì dé le yādàn," the word "yādàn" here means zero rather than literally a duck's egg. If computers can distinguish such ambiguities, it means computers have understood human languages at some level.

For the past thirty years, to solve the ambiguities of different tasks, we have developed many NLP systems, such as automatic classification of Chinese unknown words, Chinese word identification systems, and sentence parsers. At the same time, in order to construct the infrastructure for Chinese language processing, we have developed part-of-speech tagged corpora, treebanks, Chinese lexical databases, etc. These human-annotated data, however, are relatively limited, especially in contrast to the amount of text data on the Internet. Therefore a

Ph.D. graduation.



CKIP team members.

crucial goal of the NLP field is figuring out how to utilize abundant but unlabeled raw data from the Internet. In recent years, deep learning frameworks have proven effective in many NLP applications. One fundamental research topic of deep learning on NLP is called word embedding, which is how to gather lexical knowledge from a huge cache of unlabeled text data. This kind of approach is very different from that of the traditional Chinese lexical database. Now we are aiming to improve this word embedding process by the incorporation of prior knowledge bases, such as E-HowNet, a traditional Chinese lexical database we have been developing over the past ten years. Our expectations are that the learned word embeddings are more suitable for reasoning about relationships between entities and that their meanings can also be clearly interpreted. This process is very similar to a child's learning process.

In addition, we are focusing on conceptual processing of Chinese documents. The design of knowledge-based language processing systems utilizes statistical, linguistic, and commonsense knowledge provided by our evolving knowledge bases to parse the conceptual structures of sentences and interpret the meanings of sentences. Knowledge-based language processing systems incorporate knowledge bases to form a learning system. Thus, language processing systems increase their processing power due to enhancement of the knowledge bases. Conversely, the knowledge bases are evolving due to the automatic knowledge extraction made by language processing systems.

What are your expectations, both personally and for IIS?

Big data and artificial intelligence applications are now fully embraced by many sectors, including industry and education. In United States, big companies — such as Google, Facebook, Microsoft and IBM, and even new software-innovation companies — have invested substantial resources in the development of a new generation of data-mining and artificial intelligence systems. IIS occupies a favorable position to take advantage of this trend. For example, we have multimedia information processing technology at all levels, and have gained great achievements in information theory, social networks, and bioinformatics. IIS has an excellent opportunity to achieve ground-breaking performance. Increases in exchanges and cooperation with colleagues at IIS may lead to more interdisciplinary study and research in the future. Moreover, building upon the foundation of what IIS already has, I can develop new language analytical systems to strengthen IIS and help IIS have more visibility in the world. I also expect what I develop can contribute to both industry and society.

What are your suggestions for students who would like to engage in research in the field of information science?

For research, I think the most important thing is passion. Only passion for a specific area can drive oneself to get things done, and even to spend numerous hours and sleepless nights in the field and still enjoy it. Information research offers a wide variety of topics. It is very important to find the specific field that you are interested in, you are passionate about, and is also suitable for your personality. Second, research demands critical thinking, a skill that most Taiwanese students lack, mistaking it for criticism when it is instead having a clear idea of the use of reason and logic to rigorously analyze things, find the causes or the nature of things, and see advantages and disadvantages. While students must maintain the spirit of suspicion and criticism, their goal should be to help themselves have more systematic thinking and active learning. Critical thinking can help us to have more curiosity and creativity. If you can find out what is unusual in what is taken for granted, you might hold the key to a breakthrough.

Last but not the least, I want to emphasize the importance of the ability of fast learning. Online courses, such as those in Coursera and YouTube, have provided a great deal of high-quality content and have lowered the barrier for everyone to gain knowledge. Making good use of these resources and maintaining a positive spirit of learning are very important. Fast learning of needed knowledge is a necessary skill in modern society. We should take advantage of its easy availability.