The theme of this special lecture named “Multimodal knowledge graph and chatbot” , which has been given by Prof. CHUA TAT-SENG the KITHCT Chair Professor at the School of Computing, National University of Singapore. As we know the Prof. CHUA TAT-SENG is very famous in computer science, meanwhile, he is the chair of steering committee of ACM International conference on multimedia retrieval and multimedia modeling conference series.

In this special lecture, he firstly introduced the meaning of knowledge graph (KG), which organizes human knowledge about the world in a structured form in which real-world entities are inter-connected by multiple relations. At the same time, knowledge graph has recently become mainstream as large technology companies like Google, Facebook and Microsoft have announced their move from traditional search and data management to leveraging knowledge graph for semantic search and conversation system. After that, the meaning of chatbot has also been introduced, A chatbot (also known as a smartbots, talkbot, chatterbot, Bot, IM bot, interactive agent, Conversational interface or Artificial Conversational Entity) is a [computer program](https://en.wikipedia.org/wiki/Computer_program) or an [artificial intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence) which conducts a [conversation](https://en.wikipedia.org/wiki/Conversation) via auditory or textual methods.[[1]](https://en.wikipedia.org/wiki/Chatbot#cite_note-target-1)Such programs are often designed to convincingly simulate how a human would behave as a conversational partner, thereby passing the [Turing test](https://en.wikipedia.org/wiki/Turing_test).

While many research efforts on text-based KG have been carried out, little research has been done extending knowledge graph to multimodal information sources, along with multimodal search and chatbot applications. In general, this special lecture can be divided into 5 parts, the first one is MM knowledge graph, and the second part is visual relation extractions from video. In his talk, we discuss the fundamental research on knowledge extraction and from text to video, and also discussed the recent research to move from object recognition in video to relation triplet extraction. Then the development of dialog systems has been talked in the third part. In the fourth part, the theme named “multimodal dialog system” was discussed. Finally, the speaker gave his conclusion for us.

It is my pleasure to attend this special lecture which I am really interested in. In the first part about the fundamental research, “multi-channel data” has been introduced. There are many ways to obtain multi-channel data, and three main sources have been introduced. The first way is social media sources – live info streams, which can also be divided into 2 types. The first type is UGC which means spontaneous user-generated contents. And the second type named DGC, device – generated contents, we can easily through the meaning of words to know where the sources come from. At the same time, web is also the main sources which comes from web search engines, forums e-commerce sites. Wikipedia, freebase, DBpedia is the third main sources the data comes from. Expect for these, offline data, domain data, industry data for vertical domains, such as Fintech are also the import sources. Meanwhile, about the knowledge graph (KG), the Prof. CHUA TAT-SENG gave us two examples such as in a large scale semantic network to make us know KG well. Thus, we could know that the knowledge graph plays an increasingly important role in many tasks.

However, in our daily life, most research efforts are on Text-based knowledge graph. There still has many problems by using text-based KG, as we know, the internet is beyond pages and hypertext, it not only has words, but also has photos, videos, and audios are dominating the content on the web. Using only textual resource is limited in MM. Especially, in many applications such as multi-modal info retrieval, and visual Q & A. It is very important to transfer text-based KG to MMKG, an urgent need to build a multi modal knowledge graph which also named MMKG using massive multimodal resource on the web. And augment text-based KG by multimodal content.

Meanwhile, MM data offer multiple facets of knowledge. Images or video contain rich semantic or procedural knowledge and can reveal rich visual relations. For example, a simple picture can contains many useful information than a thousand words. Images or videos can go beyond the border of languages to generate a multilingual knowledge graph, even more, the rich visual relations extracted in images and videos can well describe procedural knowledges. Again the speaker define the meaning of MMKG, defined as a graph G = (V, E ), where V is the set of vertices (entities), and E is the set of edges. Meanwhile, V’s and E’s are represented by unimodal or multimodal features, such as, texts, images, videos, audios and so on. As a unified set of triples which includes subjects, predicate, object, where subject and object are arbitrary entities, and predicate is a relation. In this case, the speaker gave us an example in a complete MMKG. The example describes the physical world can be integrated into a large scale MMKG, and 3 main aspects has been introduced in this example. The first one is the enrich entities with multimodal content, and then we should insure the enable KG to perform high-level scene in understanding and reasoning. Finally, connect the vision and natural language.

In the fashion domain, as we know, multimodal knowledge could be required to fulfill various tasks or questions. For example, what does a moccasin look like? Or what look is better suited for which occasion? In this case, we need good MM analysis tools. So, we can choose MMKG for fashion domain, and several key problems need to be solved. Such as, lots of e-commerce and user-generated photos, where to buy? What to match? What to wear for what occasion? All these are challenging problems perplexing people every day. Thus, MMKG in this case, we need rich fashion domain knowledge. For MMKG, it is necessary to move from visual entities (Objects) to relations. There are several technical overviews of video relation detection, such as video object tracking, image object detection, video feature engineering, action recognition, time-series modeling and video object detection, relation prediction, and temporal localization.

In conclusion, the multimodal dialogue systems offer an effective way for information seeking and provide a general scheme for dialogue systems with in-depth visual understanding. And it also emphasizes domain knowledge incorporation for enhancing bot intelligence. The emerging of conversational recommendation systems made me feel interested in, and I am also looking forward to putting the reinforcement learning in the system in the future.