

Automated Multimodal Generation in Context-Sensitive Information Systems

Michelle Zhou¹

ABSTRACT

Imagine the next generation of information portals, where users are able to obtain information through an intelligent multimodal conversation that is tailored to the tasks they are performing, customized to their personal preferences, and adapted to their context and interaction devices. To realize this vision, we are building an intelligent user interaction framework that helps to bridge the gap between what users want and what a current system can provide.

Our framework encompasses two sets of core technologies: input technologies and output technologies. Our input technologies allow users to employ a combination of input modalities, including natural language and visual query, to express their information needs in context naturally and efficiently. On the other hand, our output technologies allow a system to automatically synthesize a multimedia response to a user's request, including both verbal and visual outputs, which is tailored to the user's interaction context, including the conversation flow and the user's personal interests.

In this talk, I will highlight the use of automated multimodal generation in both our input and output technologies. As part of our input technologies, I will present how we use automated multimodal generation technologies to dynamically create cross-modality confirmations during a user's input process. Specially, when a user employs one input modality like natural language to express his/her request ("*shipments containing T42p laptops*"), the system automatically creates the interactive representation of the same request in a complementary modality such as visual query. As a result, a user can easily switch the use of different modalities whenever needed in the course of interaction without losing the query context that s/he has built so far. Furthermore, cross-modality confirmations help to teach the user about the system's capability in supporting different input modalities. Besides supporting context-sensitive user input, automated multimodal generation is also the core piece of our output technologies. In this talk, I will focus on the practical issues in developing automated multimodal generation technologies for real-world applications. In particular, I will highlight our effort in developing optimization-based approaches to automated multimodal generation with a concrete example on a graph-matching approach to multimodal allocation.

¹ IBM T. J. Watson Research Center