Robust Multimodal Understanding for Interactive Systems

Michael Johnston
AT&T Labs Research

The ongoing convergence of the web with telephony, driven by technologies such as voice over IP, high-speed mobile data networks, and hand-held computers and smartphones, enables widespread deployment of multimodal interfaces which combine graphical user interfaces with natural human input modalities such as speech and pen. In order to support effective multimodal interaction, natural language processing techniques, which have typically been applied to linear sequences of speech or text, need to be extended to support integration and understanding of multimodal language distributed over multiple different simultaneous input modes.

Multimodal grammars (Johnston and Bangalore 2000) combine speech and gesture parsing, integration, and understanding all within a single formalism. Their finite-state implementation enables efficient processing of lattice input from speech and gesture recognition and mutual compensation for errors and ambiguities. However, like other approaches based on hand-crafted rules, multimodal grammars can be brittle with respect to unexpected, erroneous, or disfluent input.

In this talk, I will illustrate and evaluate the use of multimodal grammars to support spoken input combined with complex freehand pen input in the context of a multimodal conversational system, and explore a range of methods for improving their robustness. These include techniques for building effective language models for speech recognition when little or no training data is available and techniques for robust multimodal understanding that draw on classification, machine translation, and sequence edit methods.

Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006), pages 1–1.