*AI In Industry Column*

# Holistic Conversational Assistants

*Charles L. Ortiz, Jr.*

■ *This column describes work being done at Nuance Communications in developing virtual personal assistants (VPAs) that can engage in extended task-centered dialogues and that involve the coordination of many complex modules, along with conversational and collaborative support to such VPAs.*

One of the most exciting emerging areas of research in AI today involves the study and development of what I will refer to as *holistic systems*. These are AI systems that involve an end-to-end integration of many computational elements composing various stages of perception, language, cognition, and action. One, sort of canonical, example of such a holistic system, would be a cognitive robot in which these stages of computation were grounded in the real world (Ortiz 2016). Such holistic systems are to be contrasted with narrower systems that employ a much smaller number of elements (say, a natural language interface to a database) and in which design considerations are not as challenging.

As virtual personal assistants (VPAs) become more capable,

  ISSN 0738-4602

they will also increasingly require a holistic design perspective. At Nuance we are developing assistants that can engage in extended task-centered dialogues and that involve the coordination of many complex modules, including speech recognition, named-entity recognition (NER), morphological analysis, syntactic and semantic parsing, pragmatics, dialogue processing, reasoning and planning, and interoperation with external content sources on the web. One such system under development is an automotive assistant that is, additionally, connected to a suite of on-board actuators as well as perceptual sensors that can provide contextual information. Such a complex pipeline of technologies raises many challenges. Pragmatically speaking, such VPAs must necessarily involve a hybrid collection of technologies: there is no single existing technology that can support all of the system components.

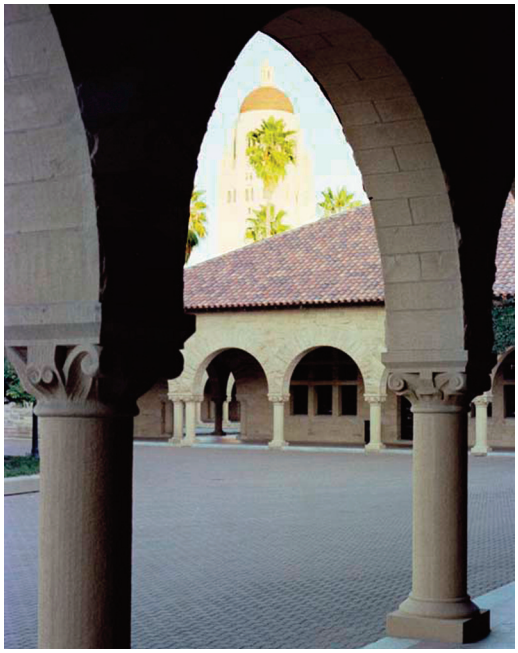
Even an individual module can involve a hybrid approach: for example, our own NL processor currently combines our own research in multicontext free grammars (Stabler 2013) with an LSTM neural net to support supertagging to speed up local parsing decisions (Lewis, Lee, and Zettlemoyer 2016). In fact, many of the system modules that we are developing involve a mixture of deep learning and symbolic approaches.

Technologically speaking, such systems present an enormous chicken and egg problem: of course, individual modules and interfaces between them must be very well defined, but attention must also be paid to the feedback paths that are necessary between component modules: in human terms, language must be integrated with cognition. For example, knowledge and reasoning should inform language processing and dialogue processing requires reasoning about tasks but must also incorporate linguistic resources. Knowledge resources must be able to make the sorts of distinctions found in language. NER can depend on information residing in back-end data stores. A natural language semantic component might make use of a linguistic ontology for the purposes of sense disambiguation, and choices at that point in the pipeline might need to be reconciled with other possibilities suggested by later reasoning. General commonsense reasoning might be enlisted to disambiguate the use of identical terms in an utterance, such as the difference in temporal span referenced by the simple appearance of the word "tonight" in "Book me a restaurant for tonight" versus "Find me a flight for tonight." Speech recognition can be improved by applying linguistic resources commonly restricted to only the natural language stage of processing. Components simply cannot be developed in isolation of each other and then patched together, and research in a technical domain cannot be conducted without consideration of the constraints imposed by other component AI technologies.

VPAs also introduce an additional real-time processing requirement: a conversation can no more suffer from delays than a robot can afford excessive deliberation in a reactive situation. This places stringent demands on the back-end reasoners to support real-time dialogue processing, for example.

A second thread of our research involves the introduction of conversational and collaborative support to VPAs. In the automotive assistant domain, the need for such support derives from the very nature of human-machine interaction in the car: a driver cannot risk visual, manual, or cognitive distraction in the course of getting information or solving a particular task (such as reserving a table at a restaurant). The collaborative dialogue systems that we are developing focus on helping users complete end-to-end tasks; these systems have sufficient self-awareness to recover from failures and also drive a conversation in productive directions. Instead, dialogue systems found in systems today are usually required to follow rigid user-system exchanges that are, for the most part, predefined rather than supporting extemporaneous and improvised dialogues. Such rigid authored dialogues can contribute to frustration and reduce the acceptance of such systems; the latter can be quite costly, particularly in today's world in which AI has created such enormous (and sometimes, unrealistic) expectations. The dialogue systems that we are developing are also well-founded in theory (Grosz and Sidner 1986) while further extended to address unique challenges that arise in personal assistant dialogues involving the revision of past decisions in a dialogue (Ortiz and Shen 2014). We are also exploring extensions having to do with multitask dialogues and the inherent trade-offs between tasks that are brought up for consideration by participants in a dialogue (Yu et al. 2016). Most recently we are investigating what we call metadialogues that will, we believe, add flexibility to VPAs by supporting graceful recovery from failure: the latter representing a major obstacle to the utility of VPA's.

The backend of our pipeline makes use of what we refer to as *big knowledge* (BK): these are large repositories of commonsense knowledge that can augment domain-specific knowledge. Big knowledge together with associated reasoning is tightly integrated with linguistic and dialogue processing. Since codifying all of commonsense knowledge is a very distant AI goal, our systems are instead designed to operate in a collaborative way such that the absence of some piece of knowledge or information does not necessarily lead to catastrophic failure: through dialogue, the system can both inform a user of a problem or provide ancillary, useful information. Secondly, the BK framework supports incremental augmentation through data deconfliction: the BK repository continually grows through inputs available through learning, crowdsourcing, or manual curation (Noessner et al. 2015). In addition, since the Internet represents a valuable, continually changing source of information, we have



## ICWSM-18 Registration Opens in March!

The Twelfth International AAAI Conference on Web and Social Media will be held June 25 – 28 at Stanford University, Palo Alto, California, USA.

ICWSM-18 will include a lively program of technical talks and posters, invited presentations, and keynote talks by Elena Grewal (Airbnb), Miguel Luengo-Oroz (UN Global Pulse), and Sarita Schoenebeck (University of Michigan).

Registration information is available at the ICWSM-18 website www.icwsm.org/2018/attending/registration. The early registration deadline is April 27, and the late registration deadline is May 25. For full details about the conference program, please visit the ICWSM-18 website (icwsm.org) or write to icwsm18@aaai.org.

developed a technology called semantic routing, that enables new content sources to be easily integrated into an intelligent system: semantic routing is able to decide on the best source of information, formulate an executable plan to acquire that information, and then fuse that information to provide answers to multipart queries or requests.

Finally, we should note that the lab in which this work was done focuses on research that leads to research prototypes, components of which are then spun off to the various product divisions at Nuance. We evaluate early prototypes through controlled user studies. The typical duration of these studies is several days, involving subject recruitment from the demographic targeted by the eventual users. As the prototypes (or specific components of the prototypes) mature, we evaluate them through either A/B studies or proof-of-concept (PoC) systems with actual customers. These tests typically take several months and reach a much larger audience. In all cases, feedback, whether from controlled user studies or PoC systems, is incorporated to improve the overall system. The work that we have described has far reaching impact to all of the Nuance product divisions in a very broad way (including mobility, enterprise, and health care). For example, all of the 150 million Nuance voice-enabled vehicles shipped globally last year are targeted for such enhancements.

## References

Grosz, B., and Sidner, C. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics* 12(3): 175–204.

Lewis, M.; Lee, K.; and Zettlemoyer, L. 2016. LSTM CCG Parsing. In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL HLT 2016). Stroudsberg, PA: Association for Computational Linguistics. doi.org/10.18653/v1/N16-1026

Noessner, J.; Martin, D.; Yeh, P.; and Patel-Schneider, P. 2015. CogMap: A Cognitive Support Approach to Property and Instance Refinement. In *The Semantic Web (ISWC 2015) 14th International Semantic Web Conference*. Lecture Notes in Computer Science volume 9366. Berlin: Springer.

Ortiz, C. 2016. Why We Need a Physically Embodied Turing Test and What It Might Look Like. *AI Magazine* 37(1): 55–62. doi.org/10.1609/aimag.v37i1.2645

Ortiz, C., and Shen, J. 2014. Dynamic Intention Structures for Dialogue Processing. Paper presented at the 18th Workshop on the Semantics and Pragmatics of Dialogue (SemDial2014), Edinburgh, UK, September 1–3. (www.macs.hw.ac.uk/InteractionLab/Semdial/semdial14.pdf)

Stabler, E. 2013. Two Models of Minimalist, Incremental Syntactic Analysis. *Topics in Cognitive Science* 5(3): 611–633. doi.org/10.1111/tops.12031

Yu, P.; Shen, J.; Yeh, P.; and Williams, B. 2016. Resolving Over-Constrained Conditional Problems Using Semantically Similar Alternatives. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (IJCAI 2016). Palo Alto, CA: AAAI Press.

**Charles Ortiz** is the director of the Nuance Laboratory for Research in Artificial Intelligence and Natural Language Processing. Prior to joining Nuance Communications, he was the director of the multiagent and multirobotics research group in the AI Center at SRI International. His own past and ongoing research has spanned the areas of causation, commonsense reasoning, collaborative agents and systems, dialogue systems, and distributed robotics. Most recently he has also been involved in various proposals, including the Nuance-sponsored Winograd Schema Challenge, to develop more suitable alternatives to the Turing test. He received his undergraduate degree from the Massachusetts Institute of Technology, and his PhD from the University of Pennsylvania. He was also a postdoctoral research fellow at Harvard University and has been a visiting scholar at Stanford and an adjunct faculty member at the University of California, Berkeley.