100 CAMBRIDGE ENGLISH CENTENARY 1913–2013

Cambridge English Centenary
Symposium on Speaking Assessment

Interactive Dialogue Systems

Diane Litman, University of Pittsburgh

The field of interactive dialogue systems uses speech and language processing to enable extended human-machine conversations. In the typical pipeline architecture of most spoken dialogue systems, a speech recognition component transcribes a spoken user utterance, a natural language understanding component extracts the transcription's syntactic structure and/or meaning, a dialogue manager determines an appropriate system response, a natural language generation component maps the response to text, and a text-to-speech component produces a spoken system utterance. There are many opportunities for assessment in the context of such a dialogue system, including utterance-level assessment of what a user says and how the user says it (e.g. to guide the real-time operation of the dialogue manager), as well as dialogue-level assessment of conversational properties such as turn-taking and dialogue structure (e.g., to evaluate a user's conversational abilities).

At the utterance-level, a dialogue manager typically uses the assessments from the speech recognition and natural language understanding components, in conjunction with an internal representation of dialogue and task state, to decide what the dialogue system should do next. For example, in a finite state dialogue manager where the dialogue states correspond to system utterances, the assessments of user utterances determine the transitions between dialogue states. Note that speech and natural language processing can be used to assess the speech files and transcriptions representing the user's utterances with respect to many linguistic dimensions. For example, in a tutorial dialogue system, syntactic analysis can been used to detect grammatical errors, while semantic analysis can be used to assess meaning with respect to an expected answer at both fine (e.g., paraphrase or entailment recognition) and coarse (e.g., on-topic or off-topic recognition) grained levels of analysis. Knowledge of pragmatics can be used to assess skills such as politeness, whereas knowledge of discourse can been used to evaluate local contextual coherence. Finally, acoustic and prosodic information particular to speech can be used to assess speaking fluency, as well as pedagogically important user states such as boredom, uncertainty, and frustration.

While utterance-level assessment in an interactive spoken dialogue system may overlap with assessment tasks in non-interactive contexts, there are often differences as well as challenges in moving to a dialogue context. For example, the goal of traditional short answer scoring is to produce a numeric score that agrees with a gold-standard human score, using statistical techniques such as lexical or semantic similarity, as well as approaches based on deeper semantic processing and inference. In contrast, the goal of short answer assessment in a dialogue system is to use similar methods to assign a label corresponding to an allowable transition from the system's current dialogue state (e.g. in a tutorial dialogue system, assessing a response to a tutor's question as correct, partially correct, or wrong, in order to reach the dialogue state corresponding to the most appropriate system feedback). Second, utterances produced during dialogue are often more spontaneous and unconstrained compared to utterances produced in non-interactive contexts, making them less predictable and harder to assess on many dimensions. As a result, compared to text, assessment of speech proficiency has focused less on aspects such as semantics, discourse and pragmatics, and more on aspects such as pronunciation and fluency. Third, the interactive capabilities of dialogue systems suggest computing and using confidence or belief information as a method to better handle noisy utterance assessments. For example, a dialogue manager with state tracking can use methods from artificial intelligence such as Bayesian networks and discriminative models to maintain a belief distribution over dialogue states as the dialogue progresses. This is in contrast to a dialogue manager that simply uses the most likely utterance assessment to select the next dialogue state, discarding any information regarding the less-likely alternatives. Another approach to handling uncertainty is to trigger a system clarification when the best assessment of a user's utterance is of low confidence. Fourth, some types of user behaviors to be assessed only occur in interactive dialogue (e.g. turn-taking). Fifth, assessments for online dialogue management must be based on linguistic features that can be computed automatically and in real-time.

At the dialogue-level, assessment typically involves higher-level and contextual user abilities that require multiple utterances of the dialogue for analysis, and that reflect the fact that dialogue is a joint activity involving two or more conversational participants. For example, in a coherent dialogue, consecutive user utterances should not be isolated and unrelated to one another. Instead, user utterances should exhibit semantic and topical relationships with both the system's and the user's history of prior utterances. In addition, user utterances should be used to achieve appropriate conversational functions, such as providing an answer after a system question, or ending the dialogue with a closing rather than a greeting. Users should also be able to use linguistic devices such as referring expressions, discourse markers, prosody, etc. that are both consistent with the underlying relationships between utterances, and that are used at appropriate times during the conversation. With respect to turn-taking abilities, users should be able to both recognize when it is their turn in a dialogue, and use linguistic signals to convey to the system that they are maintaining or ending their turn. Users must also be able to effectively ground the system's utterances, making it clear what the user has actually heard and understood, generating confirmations to the system when necessary, and appropriately recovering from system misunderstandings.

One challenge in assessing user conversational abilities is that unlike many utterance assessment tasks, there is not usually a single best reference answer. Another challenge is that due to technology limitations, user conversations with computers often exhibit somewhat different characteristics (e.g. they are simpler and more constrained) than conversations with other humans. In addition, most research in the area of dialogue has focused on understanding human dialogue abilities in order to build better spoken dialogue systems, rather than to assess user behavior along conversational dimensions. However, there are approaches being developed to evaluate the quality of simulated (i.e. computer) users of a spoken dialogue system, with respect to features such as quantity of user activity, distribution of dialogue functions of user utterances, and overall success and efficiency of the interaction. Evaluation measures have similarly been developed to evaluate the quality of dialogue systems with respect to optimizing user satisfaction. Such evaluation approaches could potentially be adapted to assess the dialogues abilities of human partners from their interactions with spoken dialogue systems.