Interpretation of Partial Utterances in Virtual Human Dialogue Systems

Kenji Sagae and **David DeVault** and **David R. Traum**
Institute for Creative Technologies
University of Southern California
Marina del Rey, CA 90292, USA
{sagae,devault,traum}@ict.usc.edu

Abstract

Dialogue systems typically follow a rigid pace of interaction where the system waits until the user has finished speaking before producing a response. Interpreting user utterances before they are completed allows a system to display more sophisticated conversational behavior, such as rapid turn-taking and appropriate use of backchannels and interruptions. We demonstrate a natural language understanding approach for partial utterances, and its use in a virtual human dialogue system that can often complete a user's utterances in real time.

1 Introduction

In a typical spoken dialogue system pipeline, the results of automatic speech recognition (ASR) for each user utterance are sent to modules that perform natural language understanding (NLU) and dialogue management only after the utterance is complete. This results in a rigid and often unnatural pacing where the system must wait until the user stops speaking before trying to understand and react to user input. To achieve more flexible turn-taking with human users, for whom turn-taking and feedback at the sub-utterance level is natural, the system needs the ability to start interpretation of user utterances before they are completed.

We demonstrate an implementation of techniques we have developed for partial utterance understanding in virtual human dialogue systems (Sagae et al., 2009; DeVault et al., 2009) with the goal of equipping these systems with sophisticated conversational behavior, such as interruptions and non-verbal feedback. Our demonstration highlights the understanding of utterances before they are finished. It also includes an utterance completion capability, where a virtual human can make a strategic decision to display its understanding of an unfinished user utterance by completing the utterance itself.

The work we demonstrate here is part of a growing research area in which new technical approaches to incremental utterance processing are being developed (e.g. Schuler et al. (2009), Kruijff et al. (2007)), new possible metrics for evaluating the performance of incremental processing are being proposed (e.g. Schlangen et al. (2009)), and the advantages for dialogue system performance and usability are starting to be empirically quantified (e.g. Skantze and Schlangen (2009), Aist et al. (2007)).

2 NLU for partial utterances

In previous work (Sagae et al., 2009), we presented an approach for prediction of semantic content from partial speech recognition hypotheses, looking at length of the speech hypothesis as a general indicator of semantic accuracy in understanding. In subsequent work (DeVault et al., 2009), we incorporated additional features of real-time incremental interpretation to develop a more nuanced prediction model that can accurately identify moments of maximal understanding within individual spoken utterances. This research was conducted in the context of the SASO-EN virtual human dialogue system (Traum et al., 2008), using a corpus of approximately 4,500 utterances from user sessions. The corpus includes a recording of each original utterance, a

$$\begin{bmatrix} mood : declarative \\ sem : \begin{bmatrix} type : event \\ agent : captain - kirk \\ event : deliver \\ theme : power - generator \\ modal : \left[possibility : can \right] \\ speech - act : \left[type : offer \right] \end{bmatrix} \end{bmatrix}$$

Figure 1: AVM utterance representation.

manual transcription, and a gold-standard semantic frame, allowing us to develop and evaluate a data-driven NLU approach.

2.1 NLU in SASO-EN Virtual Humans

Our NLU module for the SASO-EN system, mxNLU (Sagae et al., 2009), is based on maximum entropy classification (Berger et al., 1996) , where we treat entire individual semantic frames as classes, and extract input features from ASR. The NLU output representation is an attribute-value matrix (AVM), where the attributes and values represent semantic information that is linked to a domain-specific ontology and task model (Figure 1). The AVMs are linearized, using a path-value notation, as seen in the NLU input-output example below:

- Utterance (speech): *we are prepared to give you guys generators for electricity downtown*
- ASR (NLU input): *we up apparently give you guys generators for a letter city don town*
- Frame (NLU output):

```
<s>.mood declarative
<s>.sem.type event
<s>.sem.agent captain-kirk
<s>.sem.event deliver
<s>.sem.theme power-generator
<s>.sem.modal.possibility can
<s>.sem.speechact.type offer
```

When mxNLU is trained on complete ASR output for approximately 3,500 utterances, and tested on a separate set of 350 complete ASR utterances, the F-score of attribute-value pairs produced by the NLU is 0.76 (0.78 precision and 0.74 recall). These figures reflect the use of ASR at run-time, and most errors are caused by incorrect speech recognition.

2.2 NLU with partial ASR results (Sagae et al., 2009)

To interpret utterances before they are complete, we use partial recognition hypotheses produced by ASR every 200 milliseconds while the user is speaking. To process these partial utterances produced by ASR, we train length-specific models for mxNLU. These models are trained using the partial ASR results we obtain by running ASR on the audio corresponding to the utterances in the training data. The NLU task is then to predict the meaning of the entire utterance based only on a (noisy) prefix of the utterance. On average, the accuracy of mxNLU on a six-word prefix of an utterance (0.74 F-score) is almost as the same as the accuracy of mxNLU on entire utterances. Approximately half of the utterances in our corpus contain more than six words, creating interesting opportunities for conversational behavior that would be impossible under a model where each utterance must be completed before it is interpreted.

2.3 Detecting points of maximal understanding (DeVault et al., 2009)

Although length-specific NLU models produce accurate results on average, more effective use of the interpretation provided by these models might be achieved if we could automatically gauge their performance on individual utterances at run-time. To that end, we have developed an approach (DeVault et al., 2009) that aims to detect those strategic points in time, as specific utterances are occurring, when the system reaches maximal understanding of the utterance, in the sense that its interpretation will not significantly improve during the rest of the utterance.

Figure 2 illustrates the incremental output of mxNLU as a user asks, *elder do you agree to move the clinic downtown?* Our ASR processes captured audio in 200ms chunks. The figure shows the partial ASR results after the ASR has processed each 200ms of audio, along with the F-score achieved by mxNLU on each of these partials. Note that the NLU F-score fluctuates somewhat as the ASR revises its incremental hypotheses about the user utterance, but generally increases over time.

For the purpose of initiating an overlapping response to a user utterance such as this one, the agent needs to be able (in the right circumstances) to make

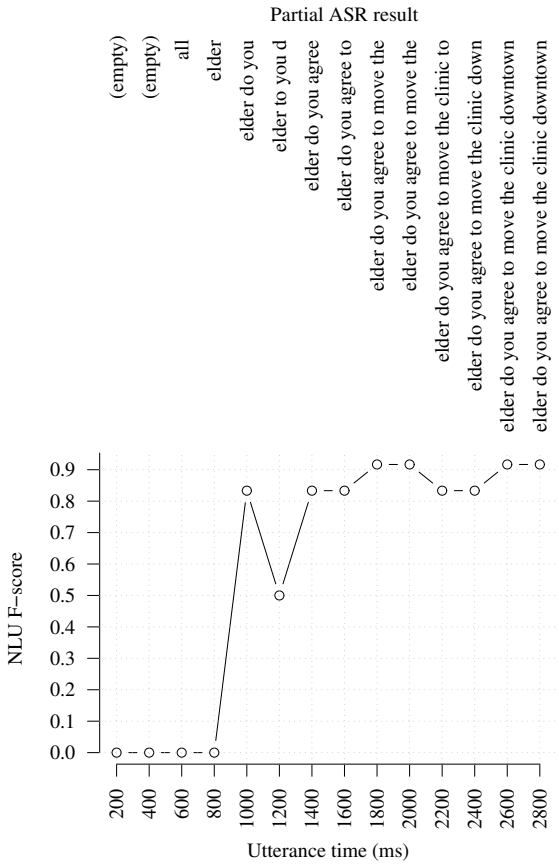

Figure 2: Incremental interpretation of a user utterance.

an assessment that it has already understood the utterance "well enough", based on the partial ASR results that are currently available. We have implemented a specific approach to this assessment which views an utterance as understood "well enough" if the agent would not understand the utterance any better than it currently does even if it were to wait for the user to finish their utterance (and for the ASR to finish interpreting the complete utterance).

Concretely, Figure 2 shows that after the entire 2800ms utterance has been processed by the ASR, mxNLU achieves an F-score of 0.91. However, in fact, mxNLU already achieves this maximal F-score at the moment it interprets the partial ASR result *elder do you agree to move the* at 1800ms. The agent therefore could, in principle, initiate an overlapping response at 1800ms without sacrificing any accuracy in its understanding of the user's utterance.

Of course the agent does not automatically realize that it has achieved a maximal F-score at 1800ms. To enable the agent to make this assessment, we have trained a classifier, which we call MAXF, that can be invoked for any specific partial ASR result, and which uses various features of the ASR result and the current mxNLU output to estimate whether the NLU F-score for the current partial ASR result is at least as high as the mxNLU F-score would be if the agent were to wait for the entire utterance.

To facilitate training of a MAXF classifier, we identified a range of potentially useful features that the agent could use at run-time to assess its confidence in mxNLU's output for a given partial ASR result. These features include: the number of partial results that have been received from the ASR; the length (in words) of the current partial ASR result; the entropy in the probability distribution mxNLU assigns to alternative output frames (lower entropy corresponds to a more focused distribution); the probability mxNLU assigns to the most probable output frame; and the most probable output frame.

Based on these features, we trained a decision tree to make the binary prediction that MAXF is TRUE or FALSE for each partial ASR result. DeVault et al. (2009) include a detailed evaluation and discussion of the classifier. To briefly summarize our results, the precision/recall/F-score of the trained MAXF model are 0.88/0.52/0.65 respectively. The high precision means that 88% of the time that the model predicts that F-score is maximized at a specific partial, it really is. Our demonstration, which we outline in the next section, highlights the utility of a high-precision MAXF classifier in making the decision whether to complete a user's utterance.

3 Demo script outline

We have implemented the approach for partial utterance understanding described above in the SASO-EN system (Traum et al., 2008), a virtual human dialogue system with speech input and output (Figure 3), allowing us to demonstrate both partial utterance understanding and some of the specific behaviors made possible by this capability. We divide this demonstration in two parts: visualization of NLU for partial utterances and user utterance completion.

Figure 3: SASO-EN: Dr. Perez and Elder al-Hassan.

Partial ASR result	*Predicted completion*
we can provide transportation	to move the patient there
the market is not	safe
there are supplies	where we are going

Table 1: Examples of user utterance completions.

3.1 Visualization of NLU for partial utterances

Because the demonstration depends on usage of the system within the domain for which it was designed, the demo operator provides a brief description of the system, task and domain. The demo operator (or a volunteer user) then speaks normally to the system, while a separate window visualizes the system's evolving understanding. This display is updated every 200 milliseconds, allowing attendees to see partial utterance understanding in action. For ease of comprehension, the display will summarize the NLU state using an English paraphrase of the predicted meaning (rather than displaying the structured frame that is the actual output of NLU). The display will also visualize the TRUE or FALSE state of the MAXF classifier, highlighting the moment the system thinks it reaches maximal understanding.

3.2 User utterance completion

The demo operator (or volunteer user) starts to speak and pauses briefly in mid-utterance, at which point, if possible, one of the virtual humans jumps in and completes the utterance (DeVault et al., 2009). Table 1 includes a few examples of the many utterances that can be completed by the virtual humans.

4 Conclusion

Interpretation of partial utterances, combined with a way to predict points of maximal understanding, opens exciting possibilities for more natural conversational behavior in virtual humans. This demonstration showcases the NLU approach and a sample application of the basic techniques.

Acknowledgments

The work described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

G. Aist, J. Allen, E. Campana, C. G. Gallo, S. Stoness, M. Swift, and M. K. Tanenhaus. 2007. Incremental dialogue system faster than and preferred to its non-incremental counterpart. In *Proc. of the 29th Annual Conference of the Cognitive Science Society*.

A. Berger, S. Della Pietra, and V. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

D. DeVault, K. Sagae, and D. Traum. 2009. Can I finish? Learning when to respond to incremental interpretation results in interactive dialogue. In *Proc. SIGDIAL*.

G. J. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, and N. Hawes. 2007. Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction. In *Proc. LangRo'2007*.

K. Sagae, G. Christian, D. DeVault, and D. R. Traum. 2009. Towards natural language understanding of partial speech recognition results in dialogue systems. In Short Paper Proceedings of NAACL HLT.

D. Schlangen, T. Baumann, and M. Atterer. 2009. Incremental reference resolution: The task, metrics for evaluation, and a Bayesian filtering model that is sensitive to disfluencies. In *Proc. SIGDIAL*, page 30–37.

W. Schuler, S. Wu, and L. Schwartz. 2009. A framework for fast incremental interpretation during speech decoding. *Computational Linguistics*, 35(3):313–343.

G. Skantze and D. Schlangen. 2009. Incremental dialogue processing in a micro-domain. In *Proc. EACL*.

D. Traum, S. Marsella, J. Gratch, J. Lee, and A. Hartholt. 2008. Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *Proc. of the Eighth International Conference on Intelligent Virtual Agents*.