Detecting the Status of a Predictive Incremental Speech Understanding Model for Real-Time Decision-Making in a Spoken Dialogue System

David DeVault, Kenji Sagae, David Traum

Institute for Creative Technologies, University of Southern California,
12015 Waterfront Drive, Playa Vista, CA 90094

{devault,sagae,traum}@ict.usc.edu

Abstract

We explore the potential for a responsive spoken dialogue system to use the real-time status of an incremental speech understanding model to guide its incremental decision-making about how to respond to a user utterance that is still in progress. Spoken dialogue systems have a range of potentially useful real-time response options as a user is speaking, such as providing acknowledgments or backchannels, interrupting the user to ask a clarification question or to initiate the system's response, or even completing the user's utterance at appropriate moments. However, implementing such incremental response capabilities seems to require that a system be able to assess its own level of understanding incrementally, so that an appropriate response can be selected at each moment. In this paper, we use a data-driven classification approach to explore the trade-offs that a virtual human dialogue system faces in reliably identifying how its understanding is progressing during a user utterance.

Index Terms: incremental speech processing, natural language understanding, spoken dialogue systems

1. Introduction

In this paper, we explore the potential for a virtual human dialogue system to use the real-time status of its incremental speech understanding model to guide its incremental decision-making about how to respond to a user utterance that is still in progress. A range of recent work has been investigating incremental speech understanding and response capabilities for spoken dialogue systems; see, e.g. [1, 2, 3, 4, 5, 6]. A general theme of this work is to relax the strict turn-taking requirements that are common in implemented systems, so that speaking with systems can become more interactive and more like speaking with a human dialogue partner. There are a range of speaker capabilities and interactive behaviors that play a role in spoken human-human dialogue, and that might be implemented, including the incremental interpretation of the speech of others, feedback on the speech of others while the speech is progressing (so-called "backchannels" [7]), monitoring of addressees and other listener feedback [8], fluent turn-taking with little or no delays [9], and overlaps of various sorts, including collaborative completions [10], repetitions and other grounding moves [11], and interruptions.

However, implementing such incremental response capabilities seems to require that a system be able to assess its own level of understanding incrementally, so that an appropriate response can be selected at each moment. For example, generating feedback in the form of a "backchannel" (such as *uh-huh*, *yeah*, *right*) during the user's speech may be interpreted by the user as implying that the system thinks it is understanding what the user is saying. Such a signal may help streamline successful communication when the system is understanding correctly, but if the system is in fact failing to understand what the user is saying, a system back-channel could lead to miscommunication and problematic consequences. Similarly, a system that is able to complete a user utterance when it thinks the utterance has been understood [3] may succeed in rapidly conveying its correct understanding, or it could cause an unnecessary repair subdialogue and diversion if its completion did not fit the user's intended meaning.

Using incremental response capabilities therefore involves trade-offs between the benefit of initiating the interactive response and the risk of doing so inappropriately. In this paper, we extend and generalize our prior work on incremental confidence estimation [3] by using a data-driven classification approach to investigate a range of different metrics that could be used by a virtual human dialogue system to assess how well its understanding is progressing during a user utterance. We then discuss the extent to which these metrics may enable the system to generate incremental response behaviors while mitigating the risks associated with doing so inappropriately.

2. Research setting

The work we present in this paper has been carried out in the setting of the SASO-EN scenario [12, 13]. We will very briefly summarize this scenario, which is designed to allow a trainee to practice multi-party negotiation skills by engaging in face to face negotiation with virtual humans. The scenario involves a negotiation about the possible re-location of a medical clinic in an Iraqi village. A human trainee plays the role of a US Army captain, and there are two virtual humans that he negotiates with: Doctor Perez, the head of an NGO clinic, and a local village elder, al-Hassan. The doctor's main objective is to treat patients. The elder's main objective is to support his village. The captain's main objective is to move the clinic out of the marketplace, ideally to the US base. Figure 1 shows the doctor and elder in the midst of a negotiation, from the perspective of the trainee.

The system has a fairly typical set of processing components for virtual humans or dialogue systems, including automatic speech recognition (ASR, mapping speech to words), natural language understanding (NLU, mapping from words to semantic frames), dialogue interpretation and management (DM, handling context, dialogue acts, reference and deciding what content to express), natural language generation (NLG, mapping frames to words), non-verbal generation, and synthesis and realization. We now turn to the design of SASO-EN's incremental NLU component, which is the focus of this paper.

28 – 31 August 2011, Florence, Italy

Figure 1: *SASO-EN negotiation in the cafe: Dr. Perez (left) looking at Elder al-Hassan.*

```
<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
```

Figure 2: *Example NLU frame.*

3. Predictive incremental speech understanding

In recent work [6, 3, 14], we have been developing a predictive incremental understanding framework for SASO-EN. We now briefly summarize this framework, to provide context for the new results reported in this paper.

Training data. The training data for our approach originates in a corpus of 3,500 utterances collected from people playing the role of captain and negotiating with the virtual doctor and elder. These user-system dialogues, which were collected with naive users, have a fairly high word error rate (average 0.39 with our current ASR configuration), with many (15%) out of domain utterances. The system is robust to these kinds of problems, both in terms of the NLU approach [6, 15] as well as the dialogue strategies [16]. This is accomplished in part by approximating the meaning of utterances.

NLU output. Utterance meanings are captured in the NLU output representation, which 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 [12]. Figure 2 shows an example representation, for an utterance such as *we can provide you with power generators*.[1] As examples of how our NLU approximates utterance meanings and is robust to some ASR errors, note that the frame in Figure 2 is also returned for an utterance of *we are prepared to give you guys generators for electricity downtown* as well as the ASR output for this utterance, *we up apparently give you guys generators for a letter city don town*. All the transcribed utterances in our corpus have been manually annotated with the correct NLU output frame.

Predictive, incremental speech understanding. Our NLU module, mxNLU [6], is based on maximum entropy classification, where we treat entire individual frames as output classes, and extract input features from partial ASR results. The specific features used by the classifier are: each word in the input string (bag-of-words representation of the input), each bigram (pairs of consecutive words), each pair of any two words in the input, and the number of words in the input string. To define the incremental understanding problem, we fed the audio of the utterances in the training data through our ASR module, which is currently PocketSphinx [17], in 200 millisecond chunks, and recorded each partial ASR result produced by the ASR. Each partial ASR result then serves as an incremental input to mxNLU, which is specially trained for partial input as discussed in [6]. NLU is predictive in the sense that, for each partial ASR result, the task of mxNLU is to produce as output the *complete* frame that has been associated by a human annotator with the user's *complete* utterance, even if that utterance has not yet been fully processed by the ASR. Our predictive framework can be contrasted with more "fully incremental" understanding frameworks, which would only try to determine the meaning of what the user has said so far at each point [2].

[1] Note that, in the figure, the hierarchical structure of the AVM has been linearized, using a path-value notation.

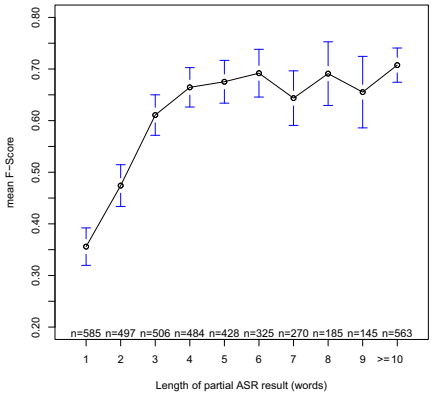

Figure 3: *Mean F-score vs. length of partial ASR result.*

Performance metric. We evaluate NLU performance by looking at precision and recall of the attribute-value pairs (or *frame elements*) that compose the predicted and correct frames for each partial ASR result. Precision represents the portion of frame elements produced by mxNLU that were correct, and recall represents the portion of frame elements in the gold-standard annotations that were proposed by mxNLU. By using precision and recall of frame elements, we take into account that certain frames are more similar than others and also allow more meaningful comparative evaluation with NLU modules that construct a frame from sub-elements or for cases when the actual frame is not in the training set. The current precision and recall of frame elements produced by mxNLU using complete ASR outputs are 0.81 and 0.75, respectively, for an F-score (harmonic mean of precision and recall) of 0.78. The performance of mxNLU for partial ASR results is discussed in detail in [6], and is summarized in Figure 3.

4. Modeling the status of incremental speech understanding

In the remainder of the paper, we explore the reliability with which various patterns in the evolution of this predictive incremental NLU's F-Score can be detected as an utterance progresses, and discuss how these patterns might be exploited in incremental response policies.

Our previous work on incremental confidence estimation

[3] has been based on the observation that mxNLU is often able to predict the correct output frame, or a partially correct output frame, before a user utterance is completed. In [3], we demonstrated that, for this corpus of utterances, it is possible for the system to detect "moments of maximal understanding" as an utterance progresses. We defined a moment of maximal understanding as follows. Given an utterance that results in a sequence of L partial ASR results, $\langle r_1, ..., r_L \rangle$, let F_t be the F-score associated with mxNLU's predicted frame at time t (based on the latest available ASR result, r_t). We then define MAXF_t for $t = 1...L$ as follows:

$$\mathrm{MAXF}_t = \begin{cases} \text{true} & \text{if } \mathrm{F_t} \geq \mathrm{F_L} \\ \text{false} & \text{otherwise} \end{cases}. \quad (1)$$

We define a moment where, in fact, $\mathrm{MAXF}_t = \text{true}$ to be a "moment of maximal understanding". (The ground truth about whether $\mathrm{MAXF}_t = \text{true}$ can be determined offline, using logged information about mxNLU's incremental outputs for the utterance, and after the utterance has been annotated and a correct output frame assigned by a human annotator.) In [3], we described the use of a machine-learning approach to build an incremental MAXF classifier, and showed that such a classifier could be constructed with a precision/recall/F-score of 0.88/0.52/0.65 respectively.

As a confidence metric, MAXF suffers from some limitations; while it predicts whether understanding will improve, it does not predict whether understanding is currently high vs. low, nor does it quantify *how much* understanding will improve. To address these limitations, we now extend and generalize this previous work by investigating the possibility of training a range of additional incremental classifiers, and using these classifiers in incremental response policies.

4.1. Metrics for assessing incremental understanding

We consider a range of new metrics for incremental speech understanding. The metrics we consider are all similar to MAXF_t, in that each one makes a binary prediction at each time t during an utterance. The metrics are defined in Table 1, and encompass a range of potentially valuable information about how well an utterance is being understood so far, and how much that understanding may improve as the user continues speaking and completes the utterance. If they could be classified reliably, each of these metrics could potentially provide valuable information for selecting an appropriate real-time response from the system. Note that in these metrics we have used an arbitrary F-Score threshold of $\frac{1}{2}$ to distinguish between "low" and "high" levels of understanding. A more optimized threshold could be used if it were available in a specific dialogue system.

4.2. Experiments and results

For each metric, we trained a decision tree using Weka's J48 training algorithm [18],[2] and using input features similar to those used to train our MAXF classifier [3]. These features included the length of the partial ASR result, the entropy in mxNLU's output distribution, the maximum probability assigned by mxNLU to any frame, the number of partial ASR results that have been returned, a unique identifier for the most probable NLU output frame, and features for each of the individual attribute-values in this output frame.

[2]Other classification models could be used as well.

Metric	Definition	Metric	Definition
High_t:	$F_t \geq \frac{1}{2}$	$\mathrm{WillBeHigh}_t$:	$F_L \geq \frac{1}{2}$
$\mathrm{Correct}_t$:	$F_t = 1$	$\mathrm{WillBeCorrect}_t$:	$F_L = 1$
$\mathrm{Incorrect}_t$:	$F_t < 1$	$\mathrm{WillBeIncorrect}_t$:	$F_L < 1$
Low_t:	$F_t < \frac{1}{2}$	$\mathrm{WillBeLow}_t$:	$F_L < \frac{1}{2}$
		MAXF_t:	$F_t \geq F_L$

Table 1: *Metrics for incremental speech understanding.*

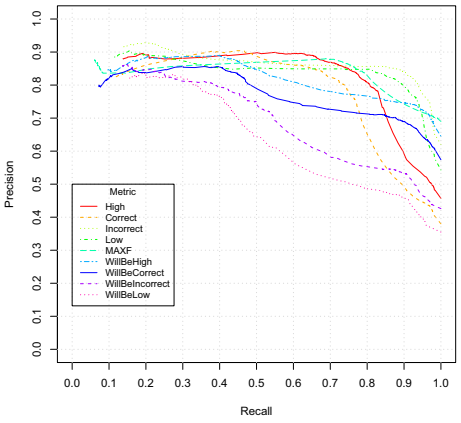

Figure 4: *Precision vs. Recall for Incremental Classification Metrics*

To assess the trained model's performance, we carried out a 10-fold cross-validation on a test corpus of 440 user utterances.[3]

For each of our trained incremental confidence classifiers, there is an associated precision-recall trade-off that describes the reliability with which the associated condition can be detected at run-time. We present these precision-recall curves in Figure 4. The results show that these trained incremental confidence models have quite different performance characteristics. For example, some metrics such as High_t, $\mathrm{Correct}_t$, Low_t, $\mathrm{Incorrect}_t$, and MAXF_t can be classified with relatively high precision for a given level of recall, while others such as $\mathrm{WillBeCorrect}_t$ and $\mathrm{WillBeIncorrect}_t$ offer a poorer precision/recall trade-off. This is perhaps intuitive, as the latter conditions may be harder to classify reliably since they depend on more precise predictions about future understanding.

As discussed above, we would like to use such incremental confidence metrics to guide our virtual humans' real-time response behaviors. For example, under some conditions, the virtual humans could give positive feedback in the form of backchannels such as head nods or saying *uh-huh* at appropriate points. However, from a design standpoint, we view it as more important for our virtual humans to not give inappropriate real-time feedback (for example, implying understanding when there is none) than it is for them to give any incremental real-time feedback. On this assumption, our observations about precision-recall trade-offs suggest that policies for generating appropriate incremental responses could perform better if framed in terms of some metrics rather than others. Consider for example these three alternative conditions under which a virtual

[3]All the partial ASR results for a given utterance were constrained to lie within the same fold, to avoid training and testing on the same utterance. Also, the test utterances were not used to train mxNLU.

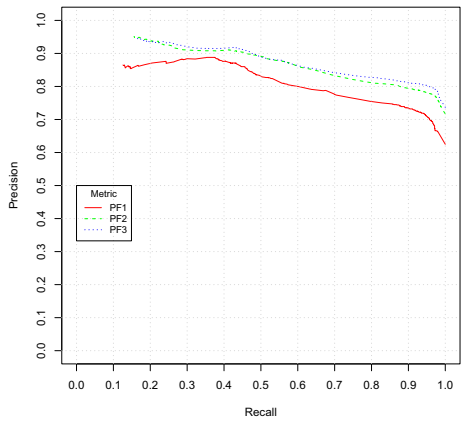

Figure 5: *Precision vs. Recall for Positive Backchannel Conditions*

human might be disposed to initiate a positive backchannel:

$$\begin{aligned} \mathrm{PF1}_t &= \mathrm{Correct}_t \vee (\mathrm{Incorrect}_t \wedge \mathrm{WillBeCorrect}_t) \\ \mathrm{PF2}_t &= \mathrm{High}_t \vee (\mathrm{Low}_t \wedge \mathrm{WillBeHigh}_t) \\ \mathrm{PF3}_t &= \mathrm{High}_t \vee (\mathrm{Low}_t \wedge \neg \mathrm{MAXF}_t) \end{aligned}$$

We might expect, based on the results in Figure 4, that $\mathrm{PF1}_t$ would be the most difficult to classify reliably, and $\mathrm{PF3}_t$ the easiest. Indeed, we trained classifiers to directly detect these three conditions (using the exact same machine-learning setup used to train classifiers for the individual metrics), and found the precision-recall trade-off depicted in Figure 5. By detecting condition $\mathrm{PF2}_t$ or $\mathrm{PF3}_t$ during incremental understanding, for example, we may be able to provide more reliable positive feedback, in the sense that these conditions can be detected with relatively high precision for a given recall level.

In future work, as we develop strategies to provide feedback based on such incremental confidence information, we expect that it will also be important to incorporate consideration of the natural timing with which backchannels and other real-time responses are provided in human-human dialogue; see e.g. [19].

5. Conclusion

In this paper, we have used a data-driven classification approach to explore a range of metrics and conditions that an implemented virtual human dialogue system could use to quantify how well its understanding is progressing during a user utterance. Using the resulting models, we have explored some of the trade-offs that this system would face in reliably providing positive feedback to users about its process of understanding. In future work, we will evaluate the run-time utility of providing this positive feedback in live interactions with users.

6. Acknowledgments

The project or effort 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.

7. References

[1] G. Skantze and A. Hjalmarsson, "Towards incremental speech generation in dialogue systems," in *Proceedings of SIGdial*, Tokyo, Japan, 2010.

[2] S. Heintze, T. Baumann, and D. Schlangen, "Comparing local and sequential models for statistical incremental natural language understanding," in *The 11th Annual Meeting of the Special Interest Group in Discourse and Dialogue (SIGDIAL 2010)*, 2010.

[3] D. DeVault, K. Sagae, and D. Traum, "Can I finish? Learning when to respond to incremental interpretation results in interactive dialogue," in *The 10th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL 2009)*, 2009.

[4] D. Schlangen and G. Skantze, "A general, abstract model of incremental dialogue processing," in *Proc. of the 12th Conference of the European Chapter of the ACL*, 2009.

[5] D. Schlangen, T. Baumann, and M. Atterer, "Incremental reference resolution: The task, metrics for evaluation, and a bayesian filtering model that is sensitive to disfluencies," in *The 10th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL 2009)*, 2009.

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

[7] V. H. Yngve, "On getting a word in edgewise," in *Papers from the Sixth Regional Meeting*. Chicago Linguistic Society, 1970, pp. 567–78.

[8] Y. I. Nakano, G. Reinstein, T. Stocky, and J. Cassell, "Towards a model of face-to-face grounding," in *ACL*, 2003, pp. 553–561.

[9] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, pp. 696–735, 1974.

[10] C. Goodwin, "The interactive construction of a sentence in natural conversation," in *Everyday Language: Studies in Ethnomethodology*, G. Psathas, Ed. New York: Ervington Press, 1979, pp. 97–121.

[11] H. H. Clark and E. F. Schaefer, "Collaborating on contributions to conversation," *Language and Cognitive Processes*, vol. 2, pp. 1–23, 1987.

[12] A. Hartholt, T. Russ, D. Traum, E. Hovy, and S. Robinson, "A common ground for virtual humans: Using an ontology in a natural language oriented virtual human architecture," in *Proc. of the Sixth International Language Resources and Evaluation (LREC'08)*, E. L. R. A. (ELRA), Ed., Marrakech, Morocco, may 2008.

[13] D. Traum, S. Marsella, J. Gratch, J. Lee, and A. Hartholt, "Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents," in *Proc. of Intelligent Virtual Agents Conference IVA-2008*, 2008.

[14] D. DeVault, K. Sagae, and D. Traum, "Incremental interpretation and prediction of utterance meaning for interactive dialogue," *Dialogue & Discourse*, vol. 2, no. 1, 2011.

[15] A. Leuski and D. Traum, "A statistical approach for text processing in virtual humans," in *26th Army Science Conference*, 2008.

[16] D. Traum, W. Swartout, J. Gratch, and S. Marsella, "A virtual human dialogue model for non-team interaction," in *Recent Trends in Discourse and Dialogue*, L. Dybkjaer and W. Minker, Eds. Springer, 2008.

[17] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *Proceedings of ICASSP*, 2006.

[18] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

[19] L.-P. Morency, I. Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Autonomous Agents and Multi-Agent Systems*, vol. 20, January 2010.