Towards Natural Language Understanding of Partial Speech Recognition Results in Dialogue Systems

Kenji Sagae and **Gwen Christian** and **David DeVault** and **David R. Traum**
Institute for Creative Technologies, University of Southern California
13274 Fiji Way, Marina del Rey, CA 90292
{sagae,gchristian,devault,traum}@ict.usc.edu

Abstract

We investigate natural language understanding of partial speech recognition results to equip a dialogue system with incremental language processing capabilities for more realistic human-computer conversations. We show that relatively high accuracy can be achieved in understanding of spontaneous utterances before utterances are completed.

1 Introduction

Most spoken dialogue systems wait until the user stops speaking before trying to understand and react to what the user is saying. In particular, in a typical dialogue system pipeline, it is only once the user's spoken utterance is complete that the results of automatic speech recognition (ASR) are sent on to natural language understanding (NLU) and dialogue management, which then triggers generation and synthesis of the next system prompt. While this style of interaction is adequate for some applications, it enforces a rigid pacing that can be unnatural and inefficient for mixed-initiative dialogue. To achieve more flexible turn-taking with human users, for whom turn-taking and feedback at the sub-utterance level is natural and common, the system needs to engage in incremental processing, in which interpretation components are activated, and in some cases decisions are made, before the user utterance is complete.

There is a growing body of work on incremental processing in dialogue systems. Some of this work has demonstrated overall improvements in system responsiveness and user satisfaction; e.g. (Aist et al., 2007; Skantze and Schlangen, 2009). Several research groups, inspired by psycholinguistic models of human processing, have also been exploring technical frameworks that allow diverse contextual information to be brought to bear during incremental processing; e.g. (Kruijff et al., 2007; Aist et al., 2007).

While this work often assumes or suggests it is possible for systems to understand partial user utterances, this premise has generally not been given detailed quantitative study. The contribution of this paper is to demonstrate and explore quantitatively the extent to which one specific dialogue system can anticipate what an utterance means, on the basis of partial ASR results, before the utterance is complete.

2 NLU for spontaneous spoken utterances in a dialogue system

For this initial effort, we chose to look at incremental processing of natural language understanding in the SASO-EN system (Traum et al., 2008), a complex spoken dialog system for which we have a corpus of user data that includes recorded speech files that have been transcribed and annotated with a semantic representation. The domain of this system is a negotiation scenario involving the location of a medical clinic in a foreign country. The system is intended as a negotiation training tool, where users learn about negotiation tactics in the context of the culture and social norms of a particular community.

2.1 The natural language understanding task

The NLU module must take the output of ASR as input, and produce domain-specific semantic frames as output. These frames are intended to capture much of the meaning of the utterance, although a

Proceedings of NAACL HLT 2009: Short Papers, pages 53–56,
Boulder, Colorado, June 2009.

dialogue manager further enriches the frame representations with pragmatic information (Traum, 2003). NLU output frames are attribute-value matrices, where the attributes and values are linked to a domain-specific ontology and task model.

Complicating the NLU task of is the relatively high word error rate (0.54) in ASR of user speech input, given conversational speech in a complex domain and an untrained broad user population.

The following example, where the user attempts to address complaints about lack of power in the proposed location for the clinic, illustrates an utterance-frame pair.

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

The original NLU component for this system was described in (Leuski and Traum, 2008). For the purposes of this experiment, we have developed a new NLU module and tested on several different data sets as described in the next section. Our approach is to use maximum entropy models (Berger et al., 1996) to learn a suitable mapping from features derived from the words in the ASR output to semantic frames. Given a set of examples of semantic frames with corresponding ASR output, a classifier should learn, for example, that when "generators" appears in the output of ASR, the value *power-generators* is likely to be present in the output frame. The specific features used by the classifier are: each word in the input string (bag-of-words representation of the input), each bigram (consecutive words), each pair of any two words in the input, and the number of words in the input string.

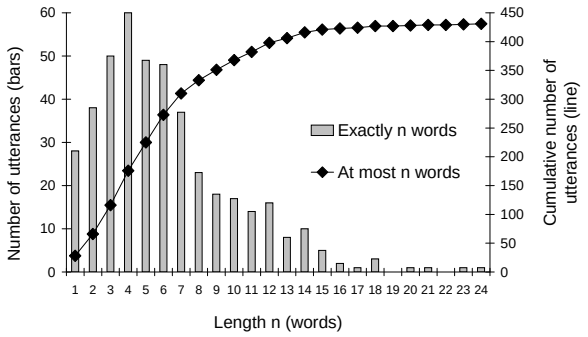

Figure 1: Length of utterances in the development set.

2.2 Data

Our corpus consists of 4,500 user utterances spread across a number of different dialogue sessions. Utterances that were out-of-domain (13.7% of the corpus) were assigned a "garbage" frame, with no semantic content. Approximately 10% of the utterances were set aside for final testing, and another 10% was designated the development corpus for the NLU module. The development and test sets were chosen so that all the utterances in a session were kept in the same set, but sessions were chosen at random for inclusion in the development and test sets.

The training set contains 136 distinct frames, each of which is composed of several attribute-value pairs, called *frame elements*. Figure 1 shows the utterance length distribution in the development set.

2.3 NLU results on complete ASR output

To evaluate NLU results, we look at precision, recall and f-score of frame elements. When the NLU module is trained on complete ASR utterances in the training set, and tested on complete ASR utterances in the development set, f-score of frame elements is 0.76, with precision at 0.78 and recall at 0.74. To gain insight on what the upperbound on the accuracy of the NLU module might be, we also trained the classifier using features extracted from gold-standard manual transcription (instead of ASR output), and tested the accuracy of analyses of gold-standard transcriptions (which would not be available at run-time in the dialogue system). Under these ideal conditions, NLU f-score is 0.87. Training on gold-standard transcriptions and testing on ASR output produces results with a lower f-score, 0.74.

3 NLU on partial ASR results

Roughly half of the utterances in our training data contain six words or more, and the average utterance length is 5.9 words. Since the ASR module is capable of sending partial results to the NLU module even before the user has finished an utterance, in principle the dialogue system can start understanding and even responding to user input as soon as enough words have been uttered to give the system some indication of what the user means, or even what the user will have said once the utterance is completed. To measure the extent to which our NLU module can predict the frame for an input utterance when it sees only a partial ASR result with the first n words, we examine two aspects of NLU with partial ASR results. The first is *correctness* of the NLU output with partial ASR results of varying lengths, if we take the gold-standard manual annotation for the entire utterance as the correct frame for any of the partial ASR results for that utterance. The second is *stability*: how similar the NLU output with partial ASR results of varying lengths is to what the NLU result would have been for the entire utterance.

3.1 Training the NLU module for analysis of partial ASR results

The simplest way to perform NLU of partial ASR results is simply to process the partial utterances using the NLU module trained on complete ASR output. However, better results may be obtained by training separate NLU models for analysis of partial utterances of different lengths. To train these separate NLU models, we first ran the audio of the utterances in the training data through our ASR module, recording all partial results for each utterance. Then, to train a model to analyze partial utterances containing n words, we used only partial utterances in the training set containing n words (unless the entire utterance contained less than n words, in which case we simply used the complete utterance). In some cases, multiple partial ASR results for a single utterance contained the same number of words, and we used the last partial result with the appropriate number of words [1]. We trained separate NLU models for n varying from one to ten.

[1] At run-time, this can be closely approximated by taking the partial utterance immediately preceding the first partial utterance of length $n + 1$.

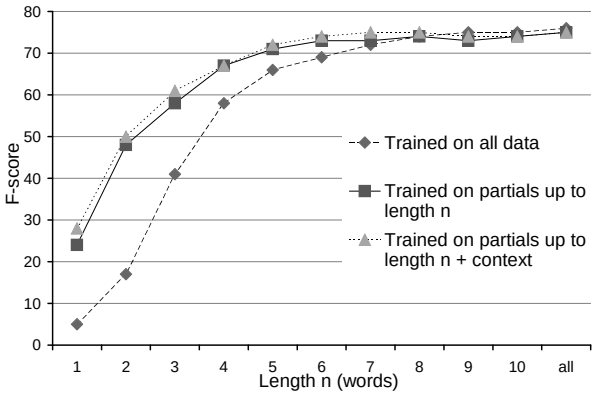

Figure 2: Correctness for three NLU models on partial ASR results up to n words.

3.2 Results

Figure 2 shows the f-score for frames obtained by processing partial ASR results up to length n using three NLU models. The dashed line is our baseline NLU model, trained on complete utterances only (model 1). The solid line shows the results obtained with length-specific NLU models (model 2), and the dotted line shows results for length-specific models that also use features that capture dialogue context (model 3). Models 1 and 2 are described in the previous sections. The additional features used in model 3 are unigram and bigram word features extracted from the most recent system utterance.

As seen in Figure 2, there is a clear benefit to training NLU models specifically tailored for partial ASR results. Training a model on partial utterances with four or five words allows for relatively high f-score of frame elements (0.67 and 0.71, respectively, compared to 0.58 and 0.66 when the same partial ASR results are analyzed using model 1). Considering that half of the utterances are expected to have more than five words (based on the length of the utterances in the training set), allowing the system to start processing user input when four or five-word partial ASR results are available provides interesting opportunities. Targeting partial results with seven words or more is less productive, since the time savings are reduced, and the gain in accuracy is modest.

The context features used in model 3 did not provide substantial benefits in NLU accuracy. It is pos-

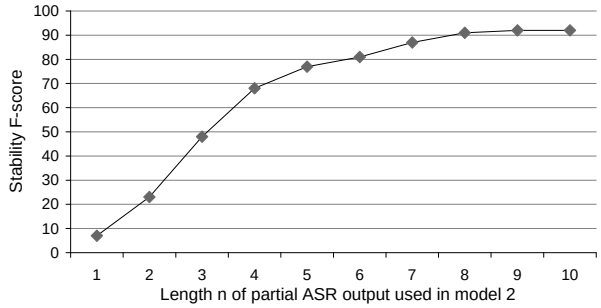

Figure 3: Stability of NLU results for partial ASR results up to length n.

sible that other ways of representing context or dialogue state may be more effective. This is an area we are currently investigating.

Finally, figure 3 shows the *stability* of NLU results produced by model 2 for partial ASR utterances of varying lengths. This is intended to be an indication of how much the frame assigned to a partial utterance differs from the ultimate NLU output for the entire utterance. This ultimate NLU output is the frame assigned by model 1 for the complete utterance. Stability is then measured as the F-score between the output of model 2 for a particular partial utterance, and the output of model 1 for the corresponding complete utterance. A stability F-score of 1.0 would mean that the frame produced for the partial utterance is identical to the frame produced for the entire utterance. Lower values indicate that the frame assigned to a partial utterance is revised significantly when the entire input is available. As expected, the frames produced by model 2 for partial utterances with at least eight words match closely the frames produced by model 1 for the complete utterances. Although the frames for partial utterances of length six are almost as accurate as the frames for the complete utterances (figure 2), figure 3 indicates that these frames are still often revised once the entire input utterance is available.

4 Conclusion

We have presented experiments that show that it is possible to obtain domain-specific semantic representations of spontaneous speech utterances with reasonable accuracy before automatic speech recognition of the utterances is completed. This allows for interesting opportunities in dialogue systems, such as agents that can interrupt the user, or even finish the user's sentence. Having an estimate of the correctness and stability of NLU results obtained with partial utterances allows the dialogue system to estimate how likely its initial interpretation of an user utterance is to be correct, or at least agree with its ultimate interpretation. We are currently working on the extensions to the NLU model that will allow for the use of different types of context features, and investigating interesting ways in which agents can take advantage of early interpretations.

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

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 *Language and Robots: Proc. from the Symposium (LangRo'2007)*. University of Aveiro, 12.

A. Leuski and D. Traum. 2008. A statistical approach for text processing in virtual humans. In *26th Army Science Conference*.

G. Skantze and D. Schlangen. 2009. Incremental dialogue processing in a micro-domain. In *Proc. of the 12th Conference of the European Chapter of the ACL*.

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 Intelligent Virtual Agents Conference IVA-2008*.

D. Traum. 2003. Semantics and pragmatics of questions and answers for dialogue agents. In *Proc. of the International Workshop on Computational Semantics*, pages 380–394, January.