Validation of a Dialog System for Language Learners

Alicia Sagae, W. Lewis Johnson, Stephen Bodnar
Alelo, Inc.
Los Angeles, CA
{asagae,ljohnson,sbodnar}@alelo.com

Abstract

In this paper we present experiments related to the validation of spoken language understanding capabilities in a language and culture training system. In this application, word-level recognition rates are insufficient to characterize how well the system serves its users. We present the results of an annotation exercise that distinguishes instances of non-recognition due to learner error from instances due to poor system coverage. These statistics give a more accurate and interesting description of system performance, showing how the system could be improved without sacrificing the instructional value of rejecting learner utterances when they are poorly formed.

1 Introduction

Conversational practice in real-time dialogs with virtual humans is a compelling element of training systems for communicative competency, helping learners acquire procedural skills in addition to declarative knowledge (Johnson, Rickel et al. 2000). Alelo's language and culture training systems allow language learners to engage in such dialogs in a serious game environment, where they practice task-based missions in new linguistic and cultural settings (Barrett and Johnson 2010). To support this capability, Alelo products apply a variety of spoken dialog technologies, including automatic speech recognition (ASR) and agent-based models of dialog that capture theories of politeness (Wang and Johnson 2008), and cultural expectations (Johnson, 2010; (Sagae, Wetzel et al. 2009).

To properly assess these dialog systems, we must take several issues into account. First, users who interact with these systems are language learners, who can be expected occasionally to produce invalid speech, and who may benefit from the corrective signal of recognizer rejection. Second, word recognition is one step in a social simulation pipeline that allows virtual humans to respond to learner input (Samtani, Valente et al. 2008). Consequently, the system goals extend beyond word-level decoding into meaning interpretation and response planning.

As a result, Word Error Rate (WER) and related metrics, such as those described by Hunt (1990) for evaluating ASR performance, are insufficient to characterize how well the speech understanding component of the dialog system performs. We need a meaningful way to account for the performance of the dialog system as a whole, which can distinguish acceptable interpretation failures from unacceptable ones.

We present a validation process for assessing speech understanding in dialog systems for language training applications. The process involves annotation of historical user data acquired from learner interaction with the Tactical Language and Culture Training System (Johnson and Valente 2009). The results indicate that learner mistakes make up the majority of non-recognitions, confirming the hypothesis that "recognition failures" are a complex category of events that are only partly explained by lack of coverage in speech understanding components such as ASR.

2 Metrics for Dialog System Assessment

Speech recognition errors in the dialog system result in at least two sub-types of error: *non-understandings*, where the system cannot find an interpretation for user input, and *misunderstandings*, where the system finds an interpretation that does not match the learner's intent (McRoy and Hirst 1995).

These classes generalize beyond speech recognition to speech understanding. This is shown in Figure 1, where "act" refers to a message

Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 241–244,
The University of Tokyo, September 24-25, 2010.

modeled along the lines of Traum & Hinkleman (1992). In the context of speech-enabled dialog systems, the understanding task is more critical, since it more closely models the overall success of the communication between the human user and the virtual human interlocutor.

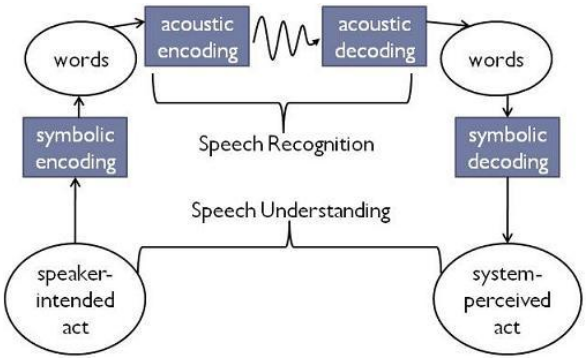

Figure 1. Speech understanding pipeline.

As a result, a variety of metrics have been suggested that assess performance at the level of intent recognition, rather than word recognition. Examples include PARADISE (Walker, Litman et al. 1998) and the work of Suendermann, Liscombe, et al (2009).

We propose an assessment procedure that uses expert annotation to compare speaker-intended acts to the acts recognized by the speech-understanding component of the dialog system. Like the metrics mentioned above, it evaluates the system's ability to recognize intent as well as words. However we focus our attention on adaptations that characterize interactions with *language learners*, who are a special type of user. As a result, we can distinguish system non-understandings and mis-understandings that are due to system error from those that are caused by learner mistakes.

Our goal is to use this information to reduce mis-understandings due to system errors; such mis-understandings can yield confusing dialog behavior, causing learners to lose confidence in the accuracy of the speech recognizer. Non-understandings may be less serious, since they occur in real life between learners and native speakers. Non-understandings due to learner error may be beneficial if the additional practice that results from non-understandings leads to an increase in language accuracy.

3 Procedure

To assess performance, we recruited two annotators to provide judgments on historical log data regarding the accuracy of the system interpretations at multiple levels, including word-level recognition and act recognition.

3.1 Annotation team and data collection

The annotators are Alelo team members with expertise in General Linguistics, French and Spanish Linguistics, Translation, and Teaching English as a Foreign Language (TEFL). Their combined experience in content authoring for Alelo courses covers more than 10 languages.

The data was collected in the fall of 2009 as part of a field test for Alelo courses teaching Iraqi Arabic and Sub-Saharan French. Naval personnel at several sites around the United States volunteered to complete the courses in self-study. The training systems generated user logs, capturing recordings of learner turns and system recognition results for each turn. From these logs, samples of beginner-level and intermediate-level dialogs were selected and anonymized for annotation.

3.2 Speech understanding accuracy

The point of this exercise is to explore how often the system fails to understand what a learner is trying to say during spoken dialog.

Annotation was performed on a total of 345 learner turns. To determine the act-level accuracy of the speech understanding system, annotators listened to the recording of each turn and selected the act they heard from a drop-down list. The results were compared with the system-perceived act result recovered from the log. Speech understanding rejections, where the system determined that no meaningful act could be perceived from the learner turn, were labeled with the act name "garbage". Human annotators could also select the garbage act for recordings where no meaningful interpretation could be made.

4 Results

To analyze the results, we measure system accuracy at two levels. First, we determine accuracy on distinguishing meaningful utterances (utterances that annotators labeled with an act) from non-meaningful speech attempts (labeled as garbage by annotators). The results are shown in Table 1. Inter-annotator agreement as measured by Cohen's Kappa on the first task is 0.8, indicating good agreement between our two experts.

Next, we examine the utterances classified as meaningful by both the system and the annota-

tors, to assess correctness at a finer level of granularity: given that the system identified the utterance as meaningful, did the meaning that it assigned match our annotators' judgments? If not, mis-understandings occur. These results are shown in Table 2. System mis-understandings over all meaningful utterances. Inter-annotator agreement on the non-understanding classification task was 0.73, suggesting that there is substantial agreement between our raters.

4.1 Correct interpretations

Numbers in the bottom-right cells of Table 1 and the first row of Table 2 represent correct system interpretations, according to an annotator. In these instances, the annotator assigned an act to the turn that matched the system interpretation for that turn (in Table 2), or both the annotator and the system assigned the label "garbage" (in Table 1). On average these examples account for 62% of the total turns.

An important result from this procedure is that it reveals the class of appropriate rejections by the speech understanding component. These "garbage-in, garbage-out" instances are instructive cases where the system indicates to the learner that he or she should re-try the utterance.

4.2 Mis-understandings

In Table 2, the row labeled "Incorrect" contains mis-understandings, where the system made an interpretation but failed to match the expert annotation. Mis-understandings account for around 3.5% of the turns in our data set, on average. The low rate of mis-understandings is an encouraging result for the overall quality of the understanding component. Prior to the introduction of the garbage model into the speech recognizer the mis-understanding rate had been relatively high, and these results indicate a significant improvement.

		Annotator 1	
		Act	Garbage
System	Act	175	3
	Garbage	94	73

		Annotator 2	
		Act	Garbage
System	Act	176	2
	Garbage	134	33

Table 1. Distinguishing meaningful utterances (corresponding to an Act) from non-meaningful attempts (Garbage).

System	**Annotator 1**	**Annotator 2**
Correct	167	160
Incorrect	8	16

Table 2. System mis-understandings over all meaningful utterances.

4.3 Non-understandings

Instances from the data set where the annotator was able to interpret an act, but the system returned "garbage," are shown in the lower-left cells of Table 1. These are system non-understandings, since the speech understanding component was not able to map the learner input to a meaningful act, even though the annotators were. Non-understandings account for 33% of turns in our data set, on average.

To understand the impact of these non-understandings on dialog system quality, we must consider the specialized case of language learners. Several components of the speech understanding pipeline are tuned with language learners in mind. For example, acoustic models used in the automatic speech recognizer are trained on a mixture of native and non-native data. The goal is for the system to be as tolerant as possible of pronunciation variability, while still catching learner mistakes.

We expect learner speech attempts to occur on a continuum, ranging from fully correct to minor mistakes to unrecoverable errors. In the first procedure, the annotators were instructed to label a recording with a meaningful act in all cases where they could do so, using garbage only for unintelligible attempts. As a result, we consciously placed the annotator tolerance at the far end of this spectrum.

Since the system is less forgiving, we hypothesize that the non-understandings we found mask two different sub-classes: instances where the system truly failed to interpret a well-formed utterance, and instances where the system was (perhaps appropriately) rejecting a learner mistake: an intelligible but malformed utterance.

In a follow-up procedure, the annotators revisited instances labeled as non-understandings. In this second round, they distinguished instances where the learner successfully performed an act that was simply outside the coverage of the speech understanding system from instances where they perceived a learner error, either in pronunciation or grammar. The results are summarized in Table 3.

We found that most of the cases of non-recognition were actually due to learner error, rather than system error.

Annotator 1	
Error Type	Count
Learner Grammar	0
Learner Pronunciation	58 (62%)
System Error	36
Total	94

Annotator 2		
Error Type	Count	κ
Learner Grammar	2	0
Learner Pronunciation	85 (63%)	0.65
System Error	47	0.65
Total	134	0.73

Table 3. Classification of non-understandings. Inter-annotator agreement (κ) is substantial over all classes.

5 Conclusions and Future Work

By applying a method for assessment that goes beyond word recognition rate, we have produced an analysis of the speech understanding components in a dialog system for language learners. Expert annotators found that most system-understood speech attempts were interpreted correctly, with mis-understandings occurring only 3% of the time. While non-understandings occurred much more frequently, a follow-up exercise showed that learner pronunciation error was the most frequent cause; these cases are legitimate candidates for system rejection, leaving 12% of all instances as non-understandings where the system was at fault. These instances represent the most beneficial errors to correct when making refinements to the speech understanding module.

In this exercise, one could interpret the human-assigned acts as a model of recognition by an extremely sympathetic hearer. Although this model may be too lenient to provide learners with realistic communication practice, it could be useful for the dialog engine to recognize some poorly-formed utterances, for the purpose of providing feedback. For example, a learner who repeatedly attempts the same utterance with unacceptable but intelligible pronunciation could trigger a tutoring-style intervention ("Are you trying to say *bonjour*? Try it more like this...").

The assessment methods and analysis presented in this paper are a first step toward this type of system improvement, one that meets the needs of language learners as a unique type of dialog-system user.

Acknowledgments

The authors thank Rebecca Row and Mickey Rosenberg for their contributions to the experiments described here, and three anonymous reviewers for comments that improved the clarity of the paper. This work was sponsored by PM TRASYS, Voice of America, the Office of Naval Research, and DARPA. Opinions expressed here are those of the author and not of the sponsors or the US Government.

References

Barrett, K. A. and W. L. Johnson (2010). Developing serious games for learning language-in-culture. Interdisciplinary Models and Tools for Serious Games: Emerging Concepts and Future Directions. R. V. Eck. Hershey, PA, IGI Global.

Hunt, M. J. (1990). "Figures of Merit for Assessing Connected Word Recognisers." Speech Communication **9**: 239-336.

Johnson, W. L., J. Rickel, et al. (2000). "Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments." Journal of Artificial Intelligence in Education **11**: 47--78.

Johnson, W. L. and A. Valente (2009). "Tactical Language and Culture Training Systems: Using AI to Teach Foreign Languages and Cultures." AI Magazine **30**(2).

McRoy, S. W. and G. Hirst (1995). "The repair of speech act misunderstandings by abductive inference." Computational Linguistics **21**(4): 435--478.

Sagae, A., B. Wetzel, et al. (2009). Culture-Driven Response Strategies for Virtual Human Behavior in Training Systems. SLaTE-2009, Warwickshire, England.

Samtani, P., A. Valente, et al. (2008). Applying the SAIBA framework to the Tactical Language and Culture Training System. AAMAS 2008 Workshop on Functional Markup Language (FML).

Suendermann, D., J. Liscombe, et al. (2009). A handsome set of metrics to measure utterance classification performance in spoken dialog systems. SigDial 2009.

Walker, M. A., D. J. Litman, et al. (1998). "Evaluating spoken dialogue agents with PARADISE: Two case studies." Computer Speech & Language **12**(4): 317-347.

Wang, N. and W. L. Johnson (2008). The Politeness Effect in an Intelligent Foreign Language Tutoring System. ITS 2008.