Learning to Balance Grounding Rationales for Dialogue Systems

Joshua Gordon
Department of Computer Science
Rebecca J. Passonneau
Center for Computational Learning Systems
Columbia University
New York, NY, USA
(joshua|becky)@cs.columbia.edu

Susan L. Epstein
Department of Computer Science
Hunter College and
The Graduate Center of the City University
of New York
New York, NY, USA
susan.epstein@hunter.cuny.edu

Abstract

This paper reports on an experiment that investigates clarification subdialogues in intentionally noisy speech recognition. The architecture learns weights for mixtures of grounding strategies from examples provided by a human wizard embedded in the system. Results indicate that the architecture learns to eliminate misunderstandings reliably despite high word error rate.

1 Introduction

We seek to develop spoken dialogue systems (*SDSs*) that communicate effectively despite uncertain input. Our thesis is that a task-oriented SDS can perform well despite a high degree of recognizer noise by relying on context. The SDS described here uses *FORRSooth*, a semi-synchronous architecture under development for task-oriented human-computer dialogue. Our immediate goals are to reduce non-understandings of user utterances (where the SDS produces no interpretation) and to eliminate *misunderstandings* (where the SDS misinterprets user utterances). The experiment recounted here investigates subdialogues consisting of an initial user response to a system prompt, and any subsequent turns that might be needed to result in full understanding of the original response. Our principal finding is that a FORRSooth-based SDS learns to build on partial understandings and to eliminate misunderstandings despite noisy ASR.

A FORRSooth-based SDS is intended to interact effectively "without the luxury of perfect components" (Paek and Horvitz, 2000), such as high-performance ASR. FORRSooth relies on portfolios of strategies for utterance interpretation and grounding, and learns to balance them from its experience. Its confidence in its interpretations is dynamically calibrated against its past experience. At each user utterance, FORRSooth selects grounding actions modulated to build upon partial interpretations in subsequent exchanges with the user.

The experiment presented here bootstraps the SDS with human expertise. In a Wizard of Oz (*WOz*) study, a person (the *wizard*) replaces selected SDS components. Knowledge is then extracted from the wizard's behavior to improve the SDS. FORRSooth uses the Relative Support Weight Learning (*RSWL*) algorithm (Epstein and Petrovic, 2006) to learn weights that balance its individual strategies. Training examples for grounding strategies are based upon examples produced by an *ablated* wizard who was restricted to the same information and actions as the system (Levin and Passonneau, 2006).

Our domain is the Andrew Heiskell Braille and Talking Book Library. Heiskell's patrons order their books by telephone, during conversation with a librarian. The next section of this paper presents related work. Subsequent sections describe the weight learning, the SDS architecture, and an experiment that challenges the robustness of utterance interpretation and grounding with intentionally noisy ASR. We

Proceedings of the SIGDIAL 2011: the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 266–271,
Portland, Oregon, June 17-18, 2011.

conclude with a discussion of the results.

2 Related Work

Despite increasingly accurate ASR methods, dialogue systems often contend with noisy ASR, which can arise from performance phenomena such as filled pauses (*er*, *um*), false starts (*fir-last name*), or noisy transmission conditions. SDSs typically experience a higher WER when deployed. For example, the WER reported for Carnegie Mellon University's Let's Go Public! went from 17% under controlled conditions to 68% in the field (Raux et al., 2005).

To limit communication errors, an SDS can rely on strategies to detect and recover from incorrect recognition output (Bohus, 2007). One such strategy, to ask the user to repeat a poorly understood utterance, can result in hyperarticulation and decreased recognition (Litman, Hirschberg and Swerts, 2006). Prior work has shown that users prefer explicit confirmation over dialogue efficiency (fewer turns) (Litman and Pan, 1999). We hypothesize that this results from an inherent tradeoff between efficiency and user confidence. We assume that evidence of partial understanding increases user confidence more than evidence of non-understanding does. FORRSooth learns to ask more questions that build on partial information, and to make fewer explicit confirmations and requests to the user to repeat herself.

While many techniques exist in the literature for semantic interpretation in task-oriented, information-seeking dialogue systems, there is no single preferred approach. SDSs rarely combine a portfolio of *NLU* (natural language understanding) resources. FORRSooth relies on "multiple processes for interpreting utterances (e.g., structured parsing versus statistical techniques)" as in (Lemon, 2003). These range from voice search (querying a database directly with ASR results) to semantic parsing.

Dialogue systems should ground their understanding of the user's objectives. To limit communication errors, an SDS can rely on strategies to detect and recover from incorrect recognition output (Bohus, 2007). In others' work, the grounding status of an utterance is typically binary (i.e., understood or not) (Allen, Ferguson and Stent, 2001; Bohus and Rudnicky, 2005; Paek and Horvitz, 2000) or ternary (i.e., understood, misunderstood, not understood) (Bohus and Rudnicky, 2009). FORRSooth's grounding decisions rely on a mixture of strategies, are based on degrees of evidence (Bohus and Rudnicky, 2009; Roque and Traum, 2009), and disambiguate among candidate interpretations. Work in (DeVault and Stone, 2009) on disambiguation in task-oriented dialogue differs from ours in that it addresses genuine ambiguities rather than noise resulting from inaccurate ASR.

3 FORR and RSWL

FORRSooth is based on *FORR* (FOr the Right Reasons), an architecture for learning and problem solving (Epstein, 1994). FORR uses sequences of decisions from multiple rationales to solve problems. Implementations have proved robust in game learning, simulated pathfinding, and constraint solving. FORR relies on an adaptive, hierarchical mixture of resource-bounded procedures called *Advisors*. Each Advisor embodies a decision rationale. Advisors' opinions (*comments*) are combined to arrive at a decision. Each comment pairs an action with a strength that indicates some degree of support for or opposition to that action. An Advisor can make multiple comments at once, and can base its comments upon descriptives. A *descriptive* is a shared data structure, computed on demand, and refreshed only when required. For each decision, FORR consults three tiers of Advisors, one tier at a time, until some tier reaches a decision.

FORR learns weights for its tier-3 Advisors with RSWL. *Relative support* is a measure of the normalized difference between the comment strength (confidence) with which an Advisor supports an action compared to other available choices. RSWL learns Advisors' weights from their comments on training examples. The degree of reinforcement (positive or negative) to an Advisor's weight is proportional to its strength and relative support for a decision.

4 FORRSooth

FORRSooth is a parallelized version of FORR. It models task-oriented dialogue with six FORR-based services that operate concurrently: INTE-

RACTION, INTERPRETATION, SATISFACTION, GROUNDING, GENERATION, and DISCOURSE. These services interpret user utterances with respect to system expectations, manage the conversational floor, and consider competing interpretations, partial understandings, and alternative courses of action. All services have access to the same data, represented by descriptives. In this section, we present background on SATISFACTION and INTERPRETATION, and provide additional detail on GROUNDING.

The role of SATISFACTION is to represent dialogue goals, and to progress towards those goals through spoken interaction. Dialogue goals are represented as *agreements*. An agreement is a subdialogue about a *target concept* (such as a specific book) whose value must be grounded through collaborative dialogue between the system and the user (Clark and Schaefer, 1989). Agreements are organized into an agreement graph that represents dependencies among them. Task-based agreements are domain specific, while grounding agreements are domain independent (cf. (Bohus, 2007)). An interpretation *hypothesis* represents the system's belief that the value of a specific target (e.g., a full name or a first name) occurred in the user's speech.

The role of INTERPRETATION is to formulate hypotheses representing the meaning of what the user has said. INTERPRETATION relies on tier-3 Advisors (essentially, mixtures of heuristics). Each Advisor constructs comments on speech recognition hypotheses. A *comment* is a semantic concept (*hypothesis*) with an associated strength. More than one Advisor can vote for the same hypothesis. Confidence in any one hypothesis is a function of votes, learned weights for Advisors, and comment strengths.

In previous work, we showed that INTERPRETATION Advisors can produce relatively reliable hypotheses given noisy ASR, with graceful degradation as recognition performance decreases (Gordon, Passonneau and Epstein, 2011). For example, at WER between 0.2 and 0.4, the concept accuracy of the top hypothesis was 80%. That work left open how to decide whether to use the top INTERPRETATION hypothesis. Here FORRSooth learns how to assess its INTERPRETATION confidence, and what grounding actions to take given different levels of confidence.

Over the life of a FORRSooth SDS, INTERPRETATION produces hypotheses for the values of target concepts. FORRSooth records the mean and variance of the comment strengths for each INTERPRETATION hypothesis, and uses them to calculate INTERPRETATION's *merit*. Merit represents FORRSooth's INTERPRETATION confidence as a dynamic, normalized estimate of the percentile in which the value falls. Merit computations improve initially with use of the SDS, and can then shift with the user population and the data. FORRSooth's approach differs from supervised confidence annotation methods that learn a fixed confidence threshold from a corpus of human-machine dialogues (Bohus, 2007).

The role of GROUNDING is to monitor the system's confidence in its interpretation of each user utterance, to provide evidence to the user of its interpretation, and to elicit corroboration, further information, or tacit agreement. To ground a target concept, FORRSooth considers one or more hypotheses for the value the user intended, and chooses a grounding action commensurate with its understanding and confidence.

GROUNDING updates the agreement graph by adding *grounding agreements* to elicit confirmations or rejections of target concepts, or to disambiguate among target concepts. A grounding agreement's *indicator target* represents the expectation of a user response. Once a sufficiently confident INTERPRETATION hypothesis is bound to an indicator target, the grounding agreement executes side effects that strengthen or weaken the hypothesis being grounded. *Recursive grounding* (where the system grounds the user's response to the system's previous grounding action) can result if the system's expectation has not been met by the next system turn.

GROUNDING makes two kinds of decisions, each with its own set of tier-3 Advisors. The first, *commit bindings*, indicates that the system is confident in the value of a target concept. In this experiment, decisions to commit to a value are irrevocable. The other kind of decision selects the next grounding utterance for any target concepts that have not yet been bound. The decision to ground a target concept is made by tier-3 Advisors that consider the distribution of hypothesis merit, as well as the success or failure of the grounding actions taken thus far.

5 FX2

FX2 is a FORRSooth SDS constructed for the current experiment. The ten FX2 INTERPRETATION Advisors are described in (Gordon, Passonneau and Epstein, 2011). Here we describe its GROUNDING actions and Advisors.

FX2 can choose among six grounding actions. Given high confidence in a single interpretation, it commits to the binding of a target value without confirmation. At slightly lower confidence levels, it chooses to implicitly confirm a target binding, with or without a hedge (e.g., the tag question "*right?*"). At even lower confidence, the grounding action is to explicitly confirm. Given competing interpretations with similarly high confidence, the grounding action is to disambiguate between the candidates. Finally, FX2 can request the user to repeat herself.

We give two examples of the twenty-three FX2 grounding Advisors. Given two interpretation hypotheses with similar confidence scores, a disambiguation Advisor votes to prompt the user to disambiguate between them. The strength for this grounding action is proportional to the ratio of the two hypotheses' scores. To avoid repeated execution of the same grounding action, one grounding Advisor votes against actions to repeat a prompt for the same target, especially if ASR confidence is low. In FX2, RSWL facilitates the use of multiple Advisors for INTERPRETATION and GROUNDING by learning weights for them that reflect their relative reliability. We describe next how we collect training examples through an ablated wizard experiment.

6 Experimental Design

This experiment tests FX2's ability to learn INTERPRETATION and GROUNDING weights. In each dialogue, FX2 introduces itself, prompts the subject for her name or a book title, and then continues the dialogue until FX2 commits to a binding for the concept, or gives up.

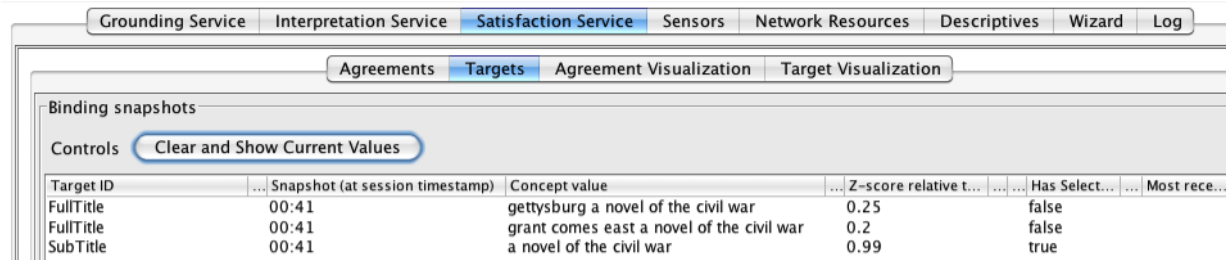
Four undergraduate native English speakers (two female, two male) participated. Speech input and output was through a microphone headset. The PocketSphinx speech recognizer produced ASR output (Huggins-Daines et al., 2006) with Wall-Street Journal dictation acoustic models adapted with ten hours of spontaneous speech. We built distinct trigram statistical language models for each type of agreement using names and titles from the Heiskell database.

We collected three data sets, referenced here as *baseline*, *wizard*, and *learning*. Each had two agreement graphs: *UserName* seeks a grounded value for the patron's full name, and *BookTitle* seeks a grounded value for a book title. 120 dialogues were collected for each dataset.

FX2 includes an optional *wizard component*. When active, the wizard component displays a GUI showing the current interpretation hypotheses for target concepts, along with their respective merit. A screen shot for the wizard GUI appears in Figure 1.

A *wizard dialogue* activates the wizard component and uses INTERPRETATION as usual, but embeds a person (the *wizard*) in GROUNDING. The wizard's purpose in this experiment is to provide training data for GROUNDING. After each user turn, the wizard makes two decisions based on data from the GUI: whether to consider any target as grounded, and which in a set of possible grounding actions to use next. The GUI displays what FX2 would choose for each decision; the wizard can either accept or override it.

Ordinarily, a FORR-based system begins with uniform Advisor weights and learns more appropriate values during its experience. Because correct interpretation and grounding are difficult tasks, however, we chose here to *prime* these weights and hypothesis merits using training examples collected during development. Development data for INTERPRETATION included 200 patron names, 400 book titles, and 50 indicator



Target ID	...	Snapshot (at session timestamp)	Concept value	...	Z-score relative t...	...	...	Has Select...	...	Most rece...
FullTitle		00:41	gettysburg a novel of the civil war		0.25			false		
FullTitle		00:41	grant comes east a novel of the civil war		0.2			false		
SubTitle		00:41	a novel of the civil war		0.99			true		

Figure 1. The wizard GUI displays hypotheses for a title from a user utterance.

Condition	Precision	Recall	F	Length
Baseline	0.65	0.78	0.72	4.36
Wizard	0.89	0.76	0.83	4.05
Learned	1.00	0.71	0.86	3.86

Table 1. Performance across three data sets.

concepts. ASR output for each item, along with its correct value, became a training example. Development data for GROUNDING came from 20 preliminary wizard dialogues. The development data also served to prime hypothesis merit.

Each subject had 30 dialogues with the system for the baseline dataset. For the wizard data set, FX2 used the same primed weights and merits as the baseline. The wizard's grounding actions and the target graphs on which they were based were saved as training examples. Weights for GROUNDING Advisors were learned from the development data training examples and the training examples saved from the wizard data set together before collecting the learned data set.

7 Results and Discussion

We assess system performance as follows. A *true positive (tp)* here is a dialogue that made no grounding errors and successfully grounded the root task agreement; a *false positive (fp)* made at least one grounding error (where the system entirely misunderstood the user). A *false negative (fn)* occurs when the system gives up on the task. Precision is *tp/(tp+fp)*, recall is *tp*/(*tp*+*fn*), and F is their mean. We measure WER using Levenshtein edit distance (Levenshtein, 1966). Because the audio data is not yet transcribed, we estimated average WER from the speaker's first known utterance (*n*=360). Overall estimated WER was 66% (54% male, 78% female).

An ideal system engages in dialogues that have high precision, high recall, and economical *dialogue length* (as measured by number of system turns). Table 1 reports that data. There is a significant increase in precision across the three data sets, a small corresponding decrease in recall, and an overall gain in F measure. The precision demonstrated by the system during dialogues in the learned data set is as good or better than that reported for our best embedded human wizards in full dialogue experiments (Ligorio, Epstein and Passonneau, 2010).

Condition	Conf	Disambig	Repeat	Other
Baseline	0.23	0.19	0.50	0.08
Wizard	0.09	0.50	0.35	0.06
Learned	0.15	0.52	0.32	0.01

Table 2. Distribution of grounding actions.

Table 2 shows the distribution of the system's grounding actions for the three data sets. The grounding actions in the learned data set are similar to the wizard's; clearly, RSWL learned well. Figure 2 illustrates an effective learned grounding behavior in which GROUNDING votes to abandon the current grounding strategy on turn 5 in favor of an approach designed to elicit an utterance from the user more likely to be processed correctly by the speech recognizer.

Our experiment suggests that misunderstandings due to ASR errors can be significantly reduced by a close integration between spoken language understanding resources and grounding behavior, together with the use of fine-grained confidence measures. Despite intentionally noisy ASR with an estimated average WER of 66%, FX2 did not experience a single misunderstanding during 120 clarification dialogues after it had trained with RSWL on examples provided by an embedded wizard. After training, the system's actions closely resembled the wizard's, indicating that the system's grounding strategies were sufficiently expressive to approximate the wizard's actions. FX2 accommodates a variety of independent spoken language understanding resources. It relies on RSWL and on merit, a self-normalizing estimate of the confidence percentile in which an interpretation hypothesis falls.

Turn	Utterance	ASR
1	What title would you like?	
2	Family and Friends	family .friends.
3	I have two guesses. The first is *Family and Friends*. The second is *Family Happiness*. Is it either of these?	
4	The first one	.nest. .first.
5	Let's try something else. Is the full title *Family and Friends*?	
6	Yes	yes

Figure 2. Example of learned GROUNDING behavior. The rightmost column is the top ASR hypothesis. Periods delimit unconfident words in the ASR.

References

James Allen, George Ferguson and Amanda Stent. 2001. An architecture for more realistic conversational systems. *Proc. 6th Int'l Conference on Intelligent User Interfaces*. ACM**:** 1-8.

Dan Bohus. 2007. Error awareness and recovery in conversational spoken language interfaces. Ph.D. thesis, Carnegie Mellon University, Pittsburgh,PA.

Dan Bohus and Alexander I. Rudnicky. 2005. Error handling in the RavenClaw dialog management framework. *Proc. Human Language Technology and Empirical Methods in Natural Language Processing*, ACL**:** 225-232.

Dan Bohus and Alexander I. Rudnicky. 2009. The RavenClaw dialog management framework: Architecture and systems. *Comput. Speech Lang.* 23(3): 332-361.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science* 13(2): 259 - 294.

David Devault and Matthew Stone. 2009. Learning to interpret utterances using dialogue history. *Proc. 12th Conference of the European Chapter of the Association for Computational Linguistics*. ACL**:** 184-192.

Susan L. Epstein. 1994. For the Right Reasons: The FORR Architecture for Learning in a Skill Domain. *Cognitive Science* 18(3): 479-511.

Susan L. Epstein and Smiljana Petrovic. 2006. Relative Support Weight Learning for Constraint Solving. *AAAI Workshop on Learning for Search*: 115-122.

Joshua B. Gordon, Rebecca J. Passonneau and Susan L. Epstein. 2011. Helping Agents Help Their Users Despite Imperfect Speech Recognition. *AAAI Symposium Help Me Help You: Bridging the Gaps in Human-Agent Collaboration*.

David Huggins-Daines, Mohit Kumar, Arthur Chan, Alan W. Black, Mosur Ravishankar and Alex I. Rudnicky. 2006. Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices. In *Proc. IEEE ICASSP, 2006.* 185-188.

Oliver Lemon. 2003. Managing dialogue interaction: A multi-layered approach. In *Proc. 4th SIGDial Workshop on Discourse and Dialogue*.

Vladimir Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*. 10**:** 707-710.

Esther Levin and Rebecca Passonneau. 2006. A WOz Variant with Contrastive Conditions. *In Proc. of Interspeech 2006 Satelite Workshop: Dialogue on Dialogues*.

Tiziana Ligorio, Susan L. Epstein and Rebecca J. Passonneau. 2010. Wizards' dialogue strategies to handle noisy speech recognition. *IEEE workshop on Spoken Language Technology (IEEE-SLT 2010)*. Berkeley, CA.

Diane Litman, Julia Hirschberg and Marc Swerts. 2006. Characterizing and predicting corrections in spoken dialogue systems. *Comput. Linguist.* 32(3): 417-438.

Diane J. Litman and Shimei Pan. 1999. Empirically evaluating an adaptable spoken dialogue system. *Proc. 7th Int'l Conference on User Modeling*. Springer-Verlag New York, Inc.**:** 55-64.

Tim Paek and Eric Horvitz. 2000. Conversation as action under uncertainty. *Proc. 16th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers Inc.**:** 455-464.

Rebecca J. Passonneau, Susan L. Epstein, Tiziana Ligorio, Joshua B. Gordon and Pravin Bhutada. 2010. Learning about voice search for spoken dialogue systems. *Human Language Technologies: NAACL 2010*. ACL**:** 840-848.

Antoine Raux, Brian Langner, Allan W. Black and Maxine Eskenazi. 2005. Let's Go Public! Taking a spoken dialog system to the real world. *Interspeech 2005 (Eurospeech)*. Lisbon, Portugal.

Antonio Roque and David Traum. 2009. Improving a virtual human using a model of degrees of grounding. *Proc. IJCAI-2009*. Morgan Kaufmann Publishers Inc.**:** 1537-1542.